NAVAL RESEARCH LOGISTICS QUARTERLY

DECEMBER 1978
VOL. 25, NO. 4

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics a will publish research and expository papers, including those in certain areas of mathematics, statistics, and economi relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, Ju September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printi Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. Cost individual issues may be obtained from the Superintendent of Documents.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulatio P-35 (Revised 1-74).

RETURN POLICIES FOR AN INVENTORY SYSTEM WITH POSITIVE AND NEGATIVE DEMANDS

Daniel P. Heyman

Bell Telephone Laboratories
Holmdel, New Jersey

ABSTRACT

We consider a single item inventory system with positive and negative stock fluctuations. Items can be purchased from a central stock, n items can be returned for a cost $R + rn$, and a linear inventory carrying cost is charged. It is shown that for minimizing the asymptotic cost rate when returns are a significant fraction of stock usage, a two-critical-number policy (a,b) is optimal, where b is the trigger level for returns and b − a is the return quantity. The values for $\underline{a}$ and b are found, as well as the operating characteristics of the system. We also consider the optimal return decision to make at time zero and show that it is partially determined by $\underline{a}$ and b.

1. INTRODUCTION

For some types of telephone central office equipment it has been observed that disconnections, rearrangements, and so on, of the equipment occur frequently enough that net demand (outflow minus inflow) may be negative in some time periods. This phenomenon may also arise whenever a supplier receives new and repaired items to satisfy customers, for example military depots. For these situations, inventory models that consider only positive demands are inadequate.

The negative demands cause the inventory level to increase, perhaps to undesirably large values. Thus, we are interested in finding a policy for returning inventory to another storage facility — which is the subject of this paper. We will consider the dynamic behavior of a single inventory location which has the option, at any time, of returning any part of its stock to a central warehouse.

The single inventory location model is of interest in its own right, but it is also important as a building block for multiechelon, multilocation models. In the telephone central-office application, many central offices are supplied from a common central stock, which can place orders on an external source. Clearly, the return policies followed at each central office should be jointly determined for optimal performance.

In an unpublished report [3], the author studied such a two-echelon, multilocation model and found that the jointly optimized policies for the second-echelon facilities were approximately the same as the individually optimized policies obtained from a single-location model, and so were the operating characteristics of the system, for reasonable values of the system's parameters. Furthermore, obtaining the jointly optimized policies is computationally prohibitive (except for exploratory research), so individual optimization provides the only practical method

for obtaining return policies. The two-echelon model mentioned above requires a large amount of detailed, but standard, manipulation to analyze; some numerical results obtained from it will be reported in this paper.

The traditional inventory models that assume that demands are nonnegative and do not permit stock returns are clearly not suitable for the problem considered here. The cash-balance inventory models (see, e.g., [2] and [10]) explicitly consider negative demands but also allow the stock level after a return to be negative, which is not appropriate for our problem. If the cash-balance models are modified so the stock level is always nonnegative, one can obtain optimal return policies for a single review-period model quite readily, but multiperiod problems lead to intractable dynamic programs [11]. Whisler [14] studied a stochastic inventory model for rented equipment. The decision to reduce the number of items rented in Whisler's model corresponds to the decision to reduce inventory in our model, but Whisler assumes that no orders can occur when the stock level is zero except between regularly spaced decision epochs, which is not appropriate for our problem. Our formulation allows a fixed charge for returns to be incorporated in the model with no extra work; fixed charges are very troublesome to add to Whisler's model. Heyman [4] considered the problem of setting a maximum stock level when there are negative demands. That model can be interpreted as a special case of this model where the return quantity is fixed at unity. Hoadley and Heyman [5] describe a one-period model for the problem discussed here. It has not been found possible to extend that model to a dynamic situation.

In section 2 we formulate our model and obtain some preliminary results for obtaining the optimal long-run policy. Results for completely specifying the model as a nonlinear program are given in section 3. This nonlinear program can be solved, in integers, using an unpublished fixed point formulation due to R. Saigal [12]. Some numerical examples are given in section 4. In section 5 we formulate the problem as a Markov-renewal program to study the return problem at time zero. Consequences of this formulation for the infinite horizon problems are given in section 6.

2. MODEL FORMULATION

In this section we will describe the basic structure of our model and obtain some preliminary results.

2.1 Basic Assumptions

To avoid confusion between items which are returned from the field and items which are sent to the central stock, we will call the former *disconnects* and the latter *returns*. The demands on the facility will be called *connects*, and items sent from the central stock to the facility will be called *shipments*.

We will make the following assumptions about costs:

(i) The cost of returning n items is $R + nr$.

(ii) The cost of shipping n items is ns.

(iii) There is a holding cost rate of h per item per unit time applied to each item in inventory.

A fixed cost is absent in (ii), because a warehouse activity should be preparing items for shipment as a matter of course. The holding-cost rate reflects the usual collection of warehousing expense, ad valorem taxes, tied up capital, and so forth.

We will assume that the inventory facility always receives a shipment immediately after requesting it from central stock, and that the connects and disconnects each form a Poisson process with rates μ and λ respectively. To avoid trivialities we assume $0 < \lambda, \mu < \infty$. These Poisson assumptions are necessary for the analysis given in section 3; useful approximate results can be obtained, if the processes are allowed to be arbitrary renewal processes, by the methods described in [4].

Let us temporarily assume a two-critical-number policy denoted (a,b), with a and b integers and $a < b$. The number b is the "trigger level" for making a return and a is the stock level after the return of size $b - a$ is made. In section 6 we will show that such a policy is indeed optimal. For the present we will concentrate on finding the optimal values of a and b, denoted a^* and b^* respectively, which minimize the asymptotic cost rate of the system, denoted by $C(a,b)$. Note that, by cost assumptions (ii) and (iii) and the assumption of zero lead time for shipments, only one item should be ordered at a time from central stock and then only when a connect occurs when the inventory is zero. We will let $C^* = C(a^*,b^*)$.

2.2 Obtaining the Cost Function and Some Conservation Laws

Let us examine the inventory process. A sample path of the stock level will have the general behavior shown in Figure 1. The arrows at t_4 and t_5 signify that a disconnect occurred when the stock level was $b - 1$, so the stock level rose to b and then instantly was reduced to a by a return. Note that the stock level is at b for only an instant, so $b - 1$ is the largest stock level actually observed.

During the time intervals $(t_0,t_1]$ and $(t_2,t_3]$ the stock level is zero, so connects occurring during these time intervals will have to be satisfied by shipments.

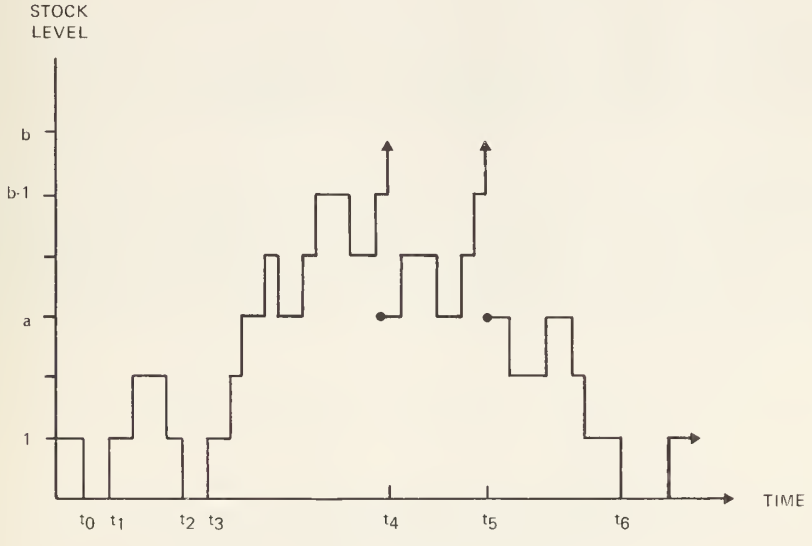

FIGURE 1. A typical realization of stock levels.

Let $\pi_0 = \pi_0(a,b)$ and $\pi_{b-1} = \pi_{b-1}(a,b)$ be the asymptotic probabilities that the stock level is zero and $b-1$ respectively. Since the connects form a Poisson process, asymptotically, the rate at which connects occur when the stock level is zero is $\mu\pi_0$. Since each such connect generates a shipment which costs s, the cost rate for shipments is $s\mu\pi_0$. Since the disconnects form a Poisson process, the rate at which returns are made is $\lambda\pi_{b-1}$. Since each return costs $R + (b-a)r$, the cost rate for returns is $[R + (b-a)r]\lambda\pi_{b-1}$. Let $L = L(a,b)$ be the average stock level in the steady state. Then hL is the asymptotic holding-cost rate. Thus, the total cost rate, $C(a,b)$, is

$$C(a,b) = s\mu\pi_0 + [R + (b-a)r]\lambda\pi_{b-1} + hL. \tag{2.1}$$

We can simplify (2.1) by establishing an equation relating π_0 and π_{b-1}. The disconnects occur at rate λ and are used either to satisfy future connect orders directly or are returned. The rate at which the former use occurs is $(1-\pi_0)\mu$, because that is the rate at which disconnects are satisfied from stock at the second echelon. The rate at which the latter use occurs is $(b-a)\lambda\pi_{b-1}$, because $\lambda\pi_{b-1}$ is the rate at which returns are made, and each return is of size $b-a$. By conservation we obtain

$$\lambda = (1-\pi_0)\mu + (b-a)(\lambda\pi_{b-1}), \tag{2.2}$$

which relates π_0 to π_{b-1}. Thus, $C(a,b)$ can be specified once π_0 and L are determined. Substituting (2.2) into (2.1) yields

$$C(a,b) = s\mu\pi_0 + [R + (b-a)r][\lambda - (1-\pi_0)\mu](b-a)^{-1} + hL. \tag{2.3}$$

Now let us consider the net demand rate (shipment rate minus return rate) a second echelon facility places on the central stock. By conservation it must be $\mu - \lambda$; if this quantity is negative, then the facility returns more than it demands in shipments, on the average.

3. THE EVALUATION OF π_0 AND L

In this section we will solve for π_0 and L; substituting these results in (2.3) we obtain $C(a,b)$.

For a particular policy (a,b), let $\pi_i = \pi_i(a,b)$ denote the steady state probability that the stock level is i, $i = 0, 1, \ldots, b-1$. Let $\rho = \lambda/\mu$; ρ is the average number of disconnects that occur between each successive pair of connects, and $(1+\rho)^{-1}$ is the probability that the next event (an *event* is a connect or a disconnect) is a connect. From standard Markov process arguments (see, e.g., [9, Chapter 2]) we know that the π_i's exist and are the unique probabilities which satisfy

$$\pi_i = (1+\rho)^{-1}\pi_{i+1} + \rho(1+\rho)^{-1}\pi_{i-1}, \quad i = 1, 2, \ldots, b-2, \ i \neq a, \tag{3.1}$$

$$\pi_0 = \rho^{-1}\pi_1, \tag{3.2}$$

$$\pi_{b-1} = \rho(1+\rho)^{-1}\pi_{b-2}, \text{ if } a \leqslant b-2, \tag{3.3}$$

$$\pi_a = (1+\rho)^{-1}\pi_{a+1} + \rho(1+\rho)^{-1}(\pi_{a-1} + \pi_{b-1}), \tag{3.4}$$

and

$$\sum_{i=0}^{b-1} \pi_i = 1. \tag{3.5}$$

When $a = b-1$ we ignore (3.3) and use (3.4) with the convention $\pi_b = 0$.

To solve these equations we first consider (3.1) as a difference equation for $i = 1,2, \ldots, a-1$ with an initial-value condition given by (3.2). Solving in the usual way, we get

$$\pi_i = \rho^i \pi_0, \quad i = 0, 1, \ldots, a. \tag{3.6}$$

Now we consider (3.1) as a difference equation for $i = a+1, a+2, \ldots, b-2$ with (3.3) as a boundary condition. The general solution to (3.1) is

$$\pi_i = c_1 + c_2 \rho^i, \tag{3.7}$$

where c_1 and c_2 are arbitrary constants. In particular,

$$\pi_{b-1} = c_1 + c_2 \rho^{b-1} \tag{3.8}$$

and

$$\pi_{b-2} = c_1 + c_2 \rho^{b-2}. \tag{3.9}$$

Substituting (3.9) into (3.3), we obtain

$$c_1 + c_2 \rho^{b-2} = \frac{1+\rho}{\rho} \pi_{b-1}. \tag{3.10}$$

Solving (3.8) and (3.10) simultaneously yields

$$c_1 = \frac{\rho}{\rho - 1} \pi_{b-1}, \quad c_2 = \frac{\pi_{b-1}}{\rho^{b-1}(1-\rho)}; \tag{3.11}$$

substituting (3.11) into (3.7) we get

$$\pi_i = \frac{\rho - \rho^{i-b+1}}{\rho - 1} \pi_{b-1}, \quad i = a+1, a+2, \ldots, b-1, \tag{3.12}$$

when $\rho \neq 1$. For $\rho = 1$, we apply L'Hospital's rule to (3.12) and obtain

$$\pi_i = \pi_{b-1}, \quad i = a+1, a+2, \ldots, b-1. \tag{3.13}$$

Substituting (3.6) and (3.12) in (3.4) we get

$$\pi_a = \frac{\rho^2(1-\rho^{a-b})}{\rho^2 - 1} \pi_{b-1} + \frac{\rho^a}{1+\rho} \pi_0. \tag{3.14}$$

From (3.6) we also have

$$\pi_a = \rho^a \pi_0. \tag{3.15}$$

From (3.14) and (3.15) we obtain

$$\pi_0 = \frac{\rho^{b-a} - 1}{\rho^{b-1}(\rho - 1)} \pi_{b-1}, \quad \rho \neq 1. \tag{3.16}$$

For $\rho = 1$,

$$\pi_0 = (b-a)\pi_{b-1}. \tag{3.17}$$

Together, (3.6), (3.12), and (3.16) express π_i in terms of π_{b-1} for $i = 0,1, \ldots, b-2$. With these and (3.5) we obtain

$$\pi_{b-1} \frac{\rho^{b-1}(1-\rho)^2}{1 - (b-a)\rho^b(1-\rho) - \rho^{b-a}}, \quad (\rho \neq 1), \tag{3.18}$$

and

$$\pi_{b-1} = \frac{2}{(b-a)(1+a+b)}, \quad (\rho = 1). \tag{3.19}$$

The steady-state mean of the process

$$L = L(a,b,\rho) = \sum_{i=1}^{b-1} i\pi_i$$

can be obtained from the appropriate equations for π_i, for example, (3.5), (3.11), (3.15) and (3.16). Hence

$$\begin{aligned} L &= \pi_o \sum_{0}^{a} i\rho^i + \pi_{b-1} \sum_{a+1}^{b-1} \frac{\rho - \rho^{i-b+1}}{\rho - 1}\, i \\ &= \pi_0 \sum_{0}^{a} i\rho^i + \frac{\pi_{b-1}}{\rho-1}\left[\sum_{a+1}^{b-1} i\rho - \rho^{-b+1} \sum_{a+1}^{b-1} i\rho^i\right] \end{aligned} \tag{3.20}$$

for $\rho \neq 1$.

The summations in (3.20) are standard geometric ones, and closed form expressions can be obtained with a reasonable amount of effort. Performing the summations yields

$$\begin{aligned} L = \pi_0 \frac{\rho}{(1-\rho)^2} \{1 - \rho^a[1 + a(1-\rho)]\} + \frac{\pi_{b-1}}{\rho-1} \\ \cdot \left[\frac{\rho}{2}(b+a)(b-a-1) - \frac{\rho^{-b+2}}{(1-\rho)^2}\right. \\ \left. \cdot \{\rho^a[a(1-\rho) + 1] - \rho^{b-1}[\rho + b(1-\rho)]\}\right] \end{aligned} \tag{3.21a}$$

for $a < b - 1$ and

$$L = \frac{\rho}{1-\rho^b} \frac{1}{1-\rho} \{1 - \rho^{b-1}[1 + (b-1)(1-\rho)]\} \tag{3.21b}$$

for $a = b - 1$. Equation (3.21b) is the steady-state expected number of customers in an $M/M/1$ queue when the queue is limited to $b-1$ customers. By collecting terms, both (3.21a) and (3.21b) can be written as

$$L = \frac{\rho}{2(1-\rho)^2}\left\{2\pi_0 - \pi_{b-1}[(1-\rho)(b^2-a^2) + (1+\rho)(b-a)]\right\}. \tag{3.22}$$

Corresponding expressions for $\rho = 1$ can be obtained in a similar manner.

From these results for π_0 and L we observe that $C(a,b)$ is a nonlinear function of a and b. Although π_0 and L depend on λ and μ only through their ratio ρ , $C(a,b)$ depends on the magnitudes of λ and μ.

The cost function can be minimized, subject to $a < b$, by a method due to Saigal [12]. This method uses a piecewise-linear approximation to the cost function and always produces an integer answer. If the approximation is convex, the answer is the global optimum; otherwise, it may not be (even if the cost function itself is convex). In section 6 we give a test for global optimality which can be used to check if the algorithm has produced the optimal solution.

4. NUMERICAL EXAMPLES

Take $R = \$10$, $r = \$2$/item, $s = \$5$/item and $h = \$2$ per item per month for an item which costs $\$100$ to purchase. Different pairs of monthly connect and disconnect rates are used. The results of some sample problems are shown in Tables 1 to 3. In these tables, the entries marked "joint" refer to optimizing the three facilities in the example jointly, including their effect on the central stock. The optimization was done by using a nonlinear programming algorithm of Saigal [12].

TABLE 1. *Sample problem I*

	Facility	Discon. Rate	Con. Rate	a	b	Prob. Empty	Prob. Full	Avg. Inv.	Cost Rate ($/mo)
Joint	1	10.5	27.5	1	50	0.618	0.0000	0.62	86.24
Ind.				4	33	0.618	0.0000	0.62	86.24
Joint	2	06.5	13.5	1	50	0.519	0.0000	0.93	36.86
Ind.				5	39	0.519	0.0000	0.93	36.86
Joint	3	03.5	4.5	4	11	0.252	0.0055	2.43	10.983
Ind.				4	11	0.258	0.0077	2.29	10.980

	Total Cost Rate	CPU (sec)
Joint	134.64	>60.
Ind.	134.64	0.48

TABLE 2. *Sample Problem II*

	Facility	Discon. Rate	Con. Rate	a	b	Prob. Empty	Prob. Full	Avg. Inv.	Cost Rate ($/mo)
Joint	1	10.5	27.5	48	50	0.618	0.0000	0.62	86.24
Ind.				4	33	0.618	0.0000	0.62	86.24
Joint	2	19.0	21.0	10	21	0.118	0.0023	5.70	25.16
Ind.				10	21	0.118	0.0023	5.70	25.16
Joint	3	21.0	19.0	5	16	0.57	0.0133	6.41	27.13
Ind.				5	16	0.57	0.0133	6.41	27.13

	Total Cost Rate	CPU (sec)
Joint	139.65	112.
Ind.	139.65	000.45

TABLE 3. *Sample Problem III*

	Facility	Discon. Rate	Con. Rate	a	b	Prob. Empty	Prob. Full	Avg. Inv.	Cost Rate ($/mo)
Joint	1	19.0	21.0	10	22	0.116	0.0019	5.86	27.17
Ind.				10	21	0.118	0.0023	5.70	27.16
Joint	2	19.0	21.0	10	22	0.116	0.0019	5.86	25.17
Ind.				10	21	0.118	0.0023	5.70	25.16
Joint	3	21.0	19.0	05	16	0.057	0.0133	6.41	27.13
Ind.				05	16	0.057	0.0133	6.41	27.13

	Total Cost Rate	CPU (sec)
Joint	78.744	122.22
Ind.	78.749	0.49

The important feature of these results is that, for these data, individual optimization does just about as well as joint optimization, and the differences in the operating characteristics are small. Furthermore, individual optimization is done in a reasonable amount of computation time and joint optimization is not.

Note that when the disconnect rate is much smaller than the connect rate, e.g., facility 1 in examples 1 and 2, two of the alternate optima for individual optimization are displayed.

5. THE TIME-ZERO RETURN PROBLEM AND A DYNAMIC PROGRAMMING FORMULATION

Now we turn to the problem of the decision to be made when the stock control system is first implemented. We call this the time-zero return problem. Since we are interested in optimizing the first decision, it is natural to study this problem with dynamic programming; in particular, we will use the structure of a Markov-renewal program [7]. Besides solving the problem at hand, this formulation will lead to further results for the infinite-horizon model studied in the previous sections.

The time-zero return problem is simply stated. Suppose at time zero the inventory level is k and the cost of returning n items at time zero is $R_0 + r_0 n$: how many items ($< k$) should be returned to minimize the long-run costs? We allow $R_0 \neq R$ and $r_0 \neq r$, but it is not necessary.

5.1 Dynamic Programming Formulation

In this section we formulate the model as a Markov-renewal decision problem. This will be the cornerstone of our analysis for the decision to be taken at time zero.

Let an *event* be a connect or disconnect. Since the connects and disconnects occur according to independent Poisson processes with rates μ and λ respectively, the time between events has a negative-exponential distribution with mean $(\lambda+\mu)^{-1}$, and the probability that the next

event is a disconnect is $\lambda/(\lambda+\mu)$. Since the negative-exponential distribution has the memoryless property, there is no loss in generality of considering the system only at event epochs. Let the *state* of the system at any time $t \geqslant 0$ be the inventory level at time t.

Let ν_i be the mean length of time the system stays in state i before changing state, p_{ij} be the probability that the next state entered is j, given that the current state is i, and d_i be the expected cost incurred during a visit to state i. To facilitate solving the model's equations it is convenient to introduce a maximum inventory level; this does not affect our previous results nor the theory behind the following results. Let $N < \infty$ be the maximum state that can be entered; we can interpret N as the size of the storage facility. Assume we use the policy (a,b) to make decisions in those states $\leqslant b$ and the decision to return nothing in those states i where $b+1 \leqslant i \leqslant N-1$, and disconnects that arrive when the inventory is at level N are freely disposed of. Then,

$$\nu_i = \begin{cases} 1/\lambda, & i=0, \\ 1/(\lambda+\mu), & i=1,2,\ \ldots,\ b-1,b+1,\ \ldots,\ N-1 \\ 0, & i=b, \\ 1/\mu, & i=N, \end{cases} \tag{5.1}$$

$$d_i = \begin{cases} \mu s/\lambda, & i=0, \\ hi/(\lambda+\mu), & i=1,2,\ \ldots,\ b-1,b+1,\ \ldots,\ N-1 \\ R+(b-a)r, & i=b, \\ hN/\mu, & i=N, \end{cases} \tag{5.2}$$

$$p_{ij} = \begin{cases} \lambda/(\lambda+\mu) \triangleq p, & j = i+1, \\ \mu/(\lambda-\mu) \triangleq q, & j = i-1, \end{cases} \tag{5.3a}$$

for $1 \leqslant i \leqslant N-1$, $i \neq b$,

$$p_{01} = p_{ba} = p_{N,N-1} = 1, \tag{5.3b}$$

and $p_{ij} = 0$ otherwise. It is clear that the policy (a,b) yields the single recurrent set of states $\{0,1,\ \ldots,b\}$ when $\lambda > 0$. When $h > 0$, then b^* can be given an *a priori* finite upper bound and we assume $b^* \leqslant N-1$. Thus, none of the pathologies of Markov-renewal programs can haunt us.

Let $v_k(t)$ be the total cost incurred by time t when the inventory at time zero is k. Jewell [7, eq. (2)] shows that, for large t,

$$v_k(t) = Ct + w_k + o(t),\ k = 0,1,\ \ldots,\ N \tag{5.4}$$

where C is the cost rate (as before, which is independent of the initial state if there is only one recurrent set of states) and $w_k = w_k(a,b)$ is the *bias* in the total cost from starting in state k. Let $w_i^* = w_i(a^*,b^*)$.

For a stationary policy (i.e., one that depends only on the state) that results in a Markov chain with transition matrix (p_{ij}) with a single finite set of recurrent states, we have [7, eq. (12)]

$$w_i + C\nu_i = d_i + \sum_j p_{ij} w_j \tag{5.5}$$

for all states i.

Now that we have formulated the model as a structured dynamic program, we will show why it is useful for our current problem.

5.2 Relation to our Current Problem

The key to the solution to our current problem is (5.4). Since the decision at time zero cannot affect the cost rate, the best one can do is affect the bias terms. If at time zero the inventory level is k and $n \geqslant 0$ items are returned, the cost incurred by time t, using the optimal policy, is

$$R_0\delta(n) + nr_0 + C^*t + w^*_{k-n} + o(t), \tag{5.6a}$$

where $\delta(n)$ is the unit step function. Expression (5.6a) is minimized by minimizing

$$f_k(n) \triangleq R_0\delta(n) + nr_0 + w^*_{k-n}, \tag{5.6b}$$

and the value of n, n^*_k say, that achieves the minimum return quantity for inventory level k at time zero is the optimal number of items to return when starting in state k. Our remaining work is to calculate the w^*_i's.

5.3 Obtaining the Bias Terms and $f_k(n)$

We shall divide our task of obtaining $f_k(n)$ into three parts. First, let us consider states zero through b inclusive. The $b + 1$ equations given by (5.5) contain $b+2$ unknowns, including C. But we can find C^* by the methods of section 2, so suppose we have done so. Now we have only $b + 1$ unknowns. Let us write our system of equations in matrix form; using an obvious notation we obtain, for any policy (a,b),

$$\underrightarrow{w} + C\underrightarrow{\nu} = \underrightarrow{d} + P\underrightarrow{w}$$

or

$$(I - P)\underrightarrow{w} = \underrightarrow{d} - C\underrightarrow{\nu}. \tag{5.7}$$

But P is a stochastic matrix, so $I - P$ has no inverse, indicating that at least one of the $b + 1$ equations is redundant. From (5.3) or by physical reasoning one can conclude that P consists of one closed class of states. From the discussion in [7, p. 955] we know there is exactly one redundant equation and any one of the equations can be discarded. We will choose the b^{th} equation as the redundant one. To reduce the number of variables by one define

$$v_i = w_i - w_0, \tag{5.8}$$

so that by replacing w_i by v_i in (5.5) and using (5.1) – (5.3) we obtain

$$v_0 = 0, \tag{5.9a}$$

$$C/\lambda = \mu s/\lambda + v_1, \tag{5.9b}$$

$$v_i + C/(\lambda+\mu) = hi/(\lambda+\mu) + pv_{i+1} + qv_{i-1}, \quad i = 1, 2, \ldots, b-1. \tag{5.9c}$$

This system has a unique solution for the relative biases v_i and (5.9c) can be used to recursively compute v_i for $i = 2, \ldots, b$.

One can obtain a closed form expression for the v_i by regarding (5.9c) as a second-order difference equation with boundary conditions (5.9a) and (5.9b). Using standard methods (see, e.g., [6, sections 1.7 and 1.8]) we obtain, for $\lambda \neq \mu$,

$$v_i = \alpha(\rho^{-i}-1) + \beta i + \gamma^2, \quad i = 0, 1, \ldots, b, \tag{5.10}$$

where

$$\alpha = \rho[(C-\mu s)/\lambda - \beta - \gamma]/(1-\rho), \tag{5.11a}$$

$$\beta = C/(\lambda-\mu) + h(\lambda+\mu)/[2(\lambda-\mu)^2], \tag{5.11b}$$

and

$$\gamma = -h/[2(\lambda-\mu)]. \tag{5.11c}$$

When $\lambda = \mu$, the solution to the difference equation can also be found by the same methods. Since (5.10) holds for any policy (a,b), it must hold for the optimal policy (a^*,b^*).

To demonstrate that the b^{th} equation is indeed redundant, we can substitute (5.10) into it and show that we obtain an identity. The b^{th} equation is

$$v_b = R + r(b-a) + v_a; \tag{5.12}$$

substituting (5.10) into (5.12) and solving for C we obtain

$$C = \mu\pi_0 s + \lambda\pi_{b-1}[R + r(b-a)] + hL \tag{5.13}$$

where π_0, π_{b-1} and L are given by (3.16), (3.18), and (3.22), respectively. Thus (5.13) above and (2.1) are identical, so (5.13) is an identity. Notice that this development emphasizes the fact that C cannot be picked arbitrarily, but that it must be given by (5.13) in order for (5.7) to be a consistent system of equations.

The second step in obtaining $f_k(n)$ is to consider those states larger than b. These states are transient, because once a state k, say, with $k \geqslant b$, is entered no state larger than b is ever revisited, and such a state k will be entered in finite time by a simple random walk analysis. Thus the return decision used in these states cannot affect the cost rate. Since we are interested in comparing the effects of different return quantities at time zero, it is fruitful to compute the bias terms for these states when the decision used in these states is to return nothing, as we have previously assumed.

Let w_i be the bias term for state i, and define

$$y_i = w_i - w_N. \tag{5.14}$$

When no returns are made in state i, using the same arguments leading to (5.9), we obtain

$$y_N = 0, \tag{5.15a}$$

$$\frac{C}{\mu} = \frac{hN}{\mu} + y_{N-1}, \tag{5.15b}$$

and

$$y_i + \frac{C}{\lambda+\mu} = \frac{hi}{\lambda+\mu} + py_{i+1} + qy_{i-1}, \quad i = b+1,b+2, \ldots, N-1. \tag{5.15c}$$

Equations (5.15) can be solved in the same way that (5.9) were solved. The difference equation solution is

$$y_i = \xi_1\rho^{N-i} + \xi_2 + \beta i + \gamma i^2, \quad i = b,b+1, \ldots, N, \tag{5.16}$$

for $\rho \neq 1$, where β and γ are as before,

$$\xi_1 = -\left\{\frac{C+hN}{\mu} + \beta + \gamma[N^2 - (N-1)^2]\right\} \Big/ (1-\rho),$$

and

$$\xi_1 = \left\{ \frac{C+hN}{\mu} + \beta[(\rho-1)N+1] + \gamma[\rho N^2 - (N-1)^2] \right\} \Big/ (1-\rho).$$

We have now obtained v_i, $i = 0, 1, \ldots, b$, and y_i, $i = b, b+1, \ldots, N$; let us now extend the domain of definition of y_i to $i = 0, 1, \ldots, b-1$. By definition $v_i = w_i - w_0$ and $y_i = w_i - w_N$ for all i, so

$$y_i - v_i = w_0 - w_N \equiv y_0. \tag{5.17}$$

In particular, (5.17) holds for $i = b$, so

$$y_0 = y_b - v_b; \tag{5.18}$$

hence y_0 can be obtained from our previous results, and so can y_i by use of (5.17).

Let us review what we have just done. First we found the bias terms *relative to* w_0 for $i = 0, 1, \ldots, b$; these are the v_i's. Next we found the bias terms *relative to* w_N for $i = b, b+1, \ldots, N$; *these are the* y_i*'s. Then we used (5.17) and (5.18) to convert the* v_i*'s into* y_i*'s for* $i = 0, 1, \ldots, b$. *Now we study* $f_k(n)$.

From (5.6b) and (5.14) we obtain

$$f_k(n) = R_0\delta(n) + nr_0 + y^*_{k-n} + w^*_N,$$

where y^*_k is y_k evaluated for the optimal policy. Even though we do not know w^*_N, it is only a constant in $f_k(n)$, hence $f_k(n)$ can be minimized over n by minimizing

$$R_0\delta(n) + nr_0 + y^*_{k-n}, \tag{5.19}$$

and doing so one will find the optimal return quantity when the inventory level is k at time zero. Once the y^*_i's have been obtained, a search of $k+1$ numbers given by (5.19) will yield the optimal return quantity. Thus we have obtained a simple algorithm to solve our problem.

When there are alternate optima for a and b, it is important to know w^*_N because it and y^*_{k-n} (and hence the optimal return quantity) will depend on the particular policy followed. Jewell [8] shows how to obtain w^*_N with simple calculations.

5.4 The Explicit Solution for an Important Special Case

An important special case of the model occurs when the movement costs are the same at all times, i.e., when $R_0 = R$ and $r_0 = r$. For this case we can partially characterize the optimal policy at time zero. Let k be the initial inventory and i^*_k be the optimal inventory level after an initial return. We define the function $\hat{f}$ by

$$\hat{f}_i = y^*_i - ri, \quad i = 0, 1, \ldots, N.$$

From (5.19) we see that i^*_k minimizes

$$R\delta(k-i) + \hat{f}_i \tag{5.20}$$

for $i = 0, 1, \ldots, N$; hence, either $i^*_k = k$ or i^*_k minimizes $\hat{f}_i$ for $i < k$. A partial characterization of i^*_k is:

THEOREM:

(I) if $k < b^*$, return nothing, i.e., $i_k^* = k$;
(II) if $k = b^*$, return down to a^*, i.e., $i_k^* = a^*$;
(III) if $k > b^*$ and $\hat{f}_i \geqslant \hat{f}_{b^*}$ for all $i > b^*$, then return down to a^*, i.e., $i_k^* = a^*$.

As usual, a^* and b^* denote the optimal infinite horizon policy, so the infinite horizon solution partially provides the solution for the returns problem at time zero.

To prove part I of this theorem, we start by observing that since (a^*,b^*) is optimal for the long run, any state i, $0 \leqslant i < b^*$, is recurrent (i.e., will be entered infinitely often) and if returning a positive amount were optimal the first time i was entered it would improve the cost rate if this amount were returned every time state i were entered (this statement is called "Howard's Improvement Routine" and is rigourously established in [13, Theorem 9.2]). But this means that (a^*,b^*) is not optimal, which is a contradiction. Thus, returning nothing must be optimal when the initial inventory is less than b^*. A similar argument proves the result for $i = b^*$, i.e., part II. Note that from (5.12) we are actually indifferent between returning $b^* - a^*$ and nothing when the initial inventory is b^*.

To prove part III, note that part I implies a^* minimizes $\hat{f}_i$ for $i \leqslant b^*$; but since $\hat{f}_i \geqslant \hat{f}_{b^*}$, a^* minimizes $\hat{f}$ for all i; hence, either $i_k^* = a^*$ or $i_k^* = k$. From (5.20) we see that, to show that a^* is optimal, we must show that $R + \hat{f}_{a^*} \leqslant \hat{f}_k$. Equation (5.12) can be written $\hat{f}_{b^*} = R + \hat{f}_{a^*}$; this, along with the hypotheses of part III, yields

$$\hat{f}_k \geqslant \hat{f}_{b^*} = R + \hat{f}_{a^*},$$

which completes the proof of part III.

6. FURTHER CONSEQUENCES OF THE DYNAMIC PROGRAMMING FORMULATION

In this section we will show how the dynamic programming formulation of section 5.1 can be used to prove the optimality of the two-critical-number policy we studied in section 3 and to provide a test of the global optimality of the solutions found by nonlinear programming.

6.1 Optimality Properties of the Two-Critical-Number Policy

We will prove that among the class of stationary policies, there is an optimal policy of the form (a,b). We start by proving that an optimal stationary policy will make returns.

PROPOSITION: There is a finite state i_c such that an optimal stationary policy insists on making a return in state i_c.

PROOF: Suppose there were no such state. Then returns would never be made and the inventory level would fluctuate exactly as the number of customers present in an $M/M/1$ queue, (see, e.g., [9] for the pertinant queueing results). From standard queueing theory it is known that if $\rho \geqslant 1$, then $L = \infty$, which is clearly not optimal because we constructed policies with a finite cost rate in section 3, so i_c exists when $\rho \geqslant 1$. For $\rho < 1$, when an optimal policy is followed, let C^* be the cost rate, T_i^* be time required for the inventory to drop from i to $i-1$, and g^* be the costs incurred by the connects and disconnects that occur while the inventrory drops from i to $i-1$. From (5.5) we obtain

$$w_i^* = hi + E(g^*) - C^* E(T_i^*) + w_{i-1}^* \tag{6.1}$$

for an optimal policy.

From the theory of the $M/M/1$ queue, $E(T_i^*)$ is the expected length of a busy period, so

$$E(T_i^*) = (\mu - \lambda)^{-1}, \quad i = 1, 2, \ldots .$$

Choose a state j so large that

$$(hj - C^*)/(\mu - \lambda) + E(g^*) > R + r. \tag{6.2}$$

From (6.1) and (6.2) we see that $w_j^* - w_{j-1}^*$ is larger under the assumed optimal policy that it would be if one item were returned. Howard's Improvement Routine then asserts that the assumed optimal policy can be improved, so that a contradiction is established, and the state j in (6.2) is an upper bound on i_c.

From this proposition it follows that if the inventory at time zero is no larger than i_c, no state larger than i_c will ever be reached. If the inventory at time zero is larger than i_c, either state i_c or a smaller one will be instantaneously reached by an initial return, or state i_c will be reached in finite time by the evolution of the inventory level when an optimal policy is followed. Let b^* be the smallest state where the optimal decision in that state is to return a positive amount, and let $b^* - a^*$ be the optimal return quantity. Then when $\mu > 0$ (as it is for any nontrivial problem) state b^* will be reached by some finite time, and thereafter no state larger than b^* will be reached. Thus, b^* and a^* will determine the minimum cost rate and the decisions in those states between $b^* + 1$, and i_c, (a^*,b^*) is optimal among all stationary policies.

This result can be strengthened to include the class of all policies when the maximum inventory level N is finite and larger than i_c. In this case the Markov-renewal program has only a finite number of states with only a finite number of choices in each state, so it has an optimal policy that is stationary [7].

6.2 Global Optimality Conditions

In section 3 we found a^* and b^* with a nonlinear programming algorithm. Since we did not prove that the function being minimized is, in general, convex, we have no guarantee that the algorithm will always find the global minimum. We can use the formulation of section 5.1 to devise a test that will determine if a given pair (a,b) is optimal.

Equation (5.13) shows that the cost rates studied in sections 3 and 5 are identical, so the solution found by the nonlinear programming algorithm is a global optimum if, and only if, it is an optimal solution to the dynamic program.

Suppose we have a policy $(\hat{a},\hat{b})$ that we think is optimal. Associated with this policy is a cost rate $\hat{C}$ and a set of relative bias terms $\hat{y}_i, i = 0, 1, \ldots, N$. Let $\nu_i^{(n)}$, $d_i^{(n)}$ and $p_{ij}^{(n)}$ be the expected duration of a visit to state i, the expected immediate cost of a visit to state i, and the conditional probability that the next state will be j, when n items are returned in state i, respectively, $i, j = 0, 1, \ldots, N$ and $n = 0, 1, \ldots, i$.

According to Jewell's Markov-renewal programming algorithm [7], the policy $(\hat{a},\hat{b})$ is optimal if, and only if,

$$\left[d_i^{(n)} + \sum p_{ij}^{(n)} \hat{y}_j - \hat{y}_i\right] \Big/ \nu_i^{(n)} \geqslant \hat{C}, \quad n = 0, 1, \ldots, i. \tag{6.3}$$

for all i.

If we take $N = 50$, (6.1) would have to be tested $\sum_{1}^{50} i = 1275$ times for each $(\hat{a},\hat{b})$ thought to be optimal, with $\hat{b} \leqslant 49$ (policies with $\hat{b} = 50$ may not be globally optimal and may just indicate that $b^* > 50$ and that the chosen bound of 50 should be increased). Each test involves only three multiplications, three additions, and one comparison once the parameters are supplied.

If a policy fails the test there are several courses of action open. If the nonlinear programming algorithm that is used is of the kind that different starting values may produce different local optima, another starting value could be tried and tested, these iterations performed until the test is passed. A second alternative is to solve the Markov-renewal program by using the algorithm in [7], for example. The linear programming formulation given below provides a third algorithm for obtaining the global optimal. Although the linear and dynamic programs are mathematically identical, the availabilty of computer codes may make one of the two formulations easier to use.

6.3 Relation to Linear Programming

It is known that Markov-renewal decision problems can be formulated as linear programs and the dynamic-programming-based algorithms are the simplex algorithms with certain block pivoting rules (see, e.g., [8]). Thus, (6.3) is just the statement that, for a vector to be optimal in an LP, all columns must price out nonnegative.

The LP formulation of our problem is

$$\text{minimize } x_{N+1}$$

$$\text{subject to } \sum_j \left(\delta_{ij} - p_{ij}^{(n)}\right) x_j + \nu_i^{(n)} x_{N+1} \geqslant d_i^{(n)}$$

for $i = 0, 1, \ldots, N$ and $n = n_i = 0, 1, \ldots, i$, where $\delta_{ij} = 1$ if $i = j$ and is zero otherwise. In the LP, x_{N+1} takes the role of the cost rate C, and x_j takes the role of the bias term w_j, $j = 0, 1, \ldots, N$. Thus, the LP contains $N + 1$ unrestricted variables and $\sum_{i=0}^{N} (i+1)$ inequality constraints.

For $N = 50$, this LP has $\sum_{1}^{51} i = 1325$ constraints. Since each of the $N + 1$ unrestricted variables has to be expressed as two nonnegative variables, and slack variables have to be added to each inquality constraint to apply the simplex method, this LP will have 1428 nonnegative variables. It does not appear that the LP formulation will work as fast as the nonlinear formulation, which has only two variables and one simple constraint. It is interesting to observe that the problem of obtaining the optimal long run policy can be formulated as either a nonlinear, linear, or dynamic program.

ACKNOWLEDGMENTS

W.C. Heller, S.B. Guthery, and especially R. Saigal, provided valuable assistance during the course of this research.

REFERENCES

[1] Cox, D.R., *Renewal Theory* (Wiley, New York, 1962).
[2] Girgis, N.M., "Optimal Cash Balance Levels," Management Science **15**, 130-140 (1968).
[3] Heyman, D.P., "A Two Echelon Inventory Model with Stock Returns," (unpublished report).
[4] Heyman, D.P., "Optimal Disposal Policies for a Single-Item Inventory System with Returns," Naval Research Logistics Quarterly **24**, 385-405 (1977).
[5] Hoadley, B., and D.P. Heyman, "A Two Echelon Inventory Model with Purchases, Dispositions, Shipments, Returns and Transhipments," Naval Research Logistics Quarterly **24**, 1-19 (1977).
[6] Hildebrand, F.B., *Finite Difference Equations and Simulations* (Prentice-Hall, Englewood Cliffs, N.J., 1968).
[7] Jewell, W.S., "Markov Renewal Programming II: Infinite Return Models," Operations Research **11**, 949-971 (1963).
[8] Jewell, W.S., "A Note on Markov-Renewal Programming," Operations Research Center ORC 66-41, University of California, Berkeley (1966).
[9] Kleinrock, L., *Queueing Systems* (Wiley, New York, 1975).
[10] Neave, E.H., "The Stochastic Cash Balance Problem with Fixed Costs for Increases and Decreases," Management Science **16**, 472-290 (1970).
[11] Rolfe, A.J., "Single Location Inventory Model with Stock Returns" (unpublished report).
[12] Saigal, R., private communication, 1975.
[13] Strauch, R.E., "Negative Dynamic Programming," Annals of Mathematical Statistics **37**, 871-890 (1966).
[14] Whisler, W.D., "A Stochastic Inventory Model for Rental Equipment," Management Science **13**, 640-647 (1967).

THE DISTRIBUTION OF RECOVERABLE INVENTORY ITEMS FROM A REPAIR CENTER WHEN THE NUMBER OF CONSUMPTION CENTERS IS LARGE*

Bruce L. Miller

System Science Department
University of California, Los Angeles
Los Angeles, California

Mohammad Modarres-Yazdi

Industrial Engineering Department
Aryamehr University
Tehran, Iran

ABSTRACT

This paper considers real-time decision rules for an inventory system where items are repaired than "used up." The problem is to decide which user in the system has the greatest need for the newly available inventory items coming out of repair. The main result shows that two published approahes, the Transportation Time Look Ahead policy and METRIC, are optimal when the number of users gets large. A useful byproduct of the proof is a lower bound on the average backorder rate for a repair-inventory system of any size.

1. INTRODUCTION AND SUMMARY

We consider a repair-inventory system where the inventory items are repaired when they fail. This contrasts in a significant way with the standard inventory model in which a failure (or demand) means that the item is used up and replaced by a new item. In the U.S. Air Force inventory system which motivates this study, well over half of the multibillion inventory value is in repairable items.

Study of a repair-inventory system consists of two problems: (1) How should the inventory items be utilized in a system of a repair center and m consumption centers (the distribution problem)? (2) How many items should be purchased (the procurement problem)? Our analysis is concerned with the distribution problem.

The model we will analyze has been investigated by several authors, and a survey of the earlier work is included in the paper of Simon [5]. The most recent work in the distribution problem is that of Miller [2] which is described in the next paragraph. The METRIC model of Sherbrooke [4] is still the standard for the procurement problem. The METRIC model is also a distribution procedure, and this paper ends with some positive results on METRIC for the distribution problem.

*This research was partially supported by the National Science Foundation under ENG 74-13494 and ENG 76-12230. Most of the results in this paper came from the Ph.D. dissertation of Mohammad Modarres-Yazdi entitled, "Myopic Decision Rules in the Optimization of Queueing and Inventory Models."

The decision-making in the distribution problem is concerned with answering the question: "To which consumption center should we ship the newly repaired item?" Our proposed solution to the distribution problem, the Transportation Time Look Ahead policy, was introduced in the RAND report [1]. It was presented as a heuristic policy which determines the "neediest" comsumption center (Air Force base), and sends the newly repaired item to that center from the repair center. This policy was tested by simulation, and it resulted on the average in 20 percent lower backorder rate than the METRIC model of Sherbrooke [4]. Furthermore, it was less computationally demanding. Calculating an optimal solution to the distribution problem is impossible, since formally it is a dynamic programming problem with many state variables, representing the inventory level at the consumption centers, the number of items en route to the consumption centers, the number of items in repair, and the number of items en route to the repair center from the consumption centers.

A theoretical justification for the Transportation Time Look Ahead policy was obtained by Miller [2]. It was shown to be optimal for a modified version of the problem whose main assumption was that there is no repair at all, that is, every failure at a consumption center is matched by a repaired item simultanerously appearing at the repair center. Physically, this is very far from reality, of course, but mathematically it represented a decision problem similar to the true system. In this paper we will show that a modified version of the Transportation Time Look Ahead policy is optimal for the true system when the number of consumption centers goes to infinity. The criterion for optimality is the average backorder rate per unit time per consumption center in an infinite horizon.

The proof of optimality is carried out by first obtaining a lower bound for the average backorder rate using an arbitrary policy. This bound does not depend on the size of the system, and therefore it is useful in itself as we can compare the backorder rate of a given policy obtained by simulation with this lower bound. The second part of the proof entails showing that the Transportation Time Look Ahead policy approaches this bound as the number of consumption centers goes to infinity. This proof is also used to show that the METRIC distribution model is optimal.

2. THE MODEL

The model will be described in general (non-Air Force) terms for a single inventory item. The system has a repair center, m consumption centers, and N spare items.

Each consumption center supports activities which produce failures of the item according to a Poisson rate λ_i of items per day. This is assumed to be the case even when the inventory shortage hinders these activities, since it is assumed that the available units (aircraft in the Air Force example) can be utilized at a higher rate to compensate for the shortage. In addition to the parameter λ_i, a second parameter is needed to describe each consumption center, namely T_i, the one-way deterministic shipping time between the consumption center and the repair center. In order to limit the detrimental effect of these inventory shortages, the system employs N spare units which are used to replace operational units when they fail.

When a unit fails at consumption center i, a spare, if available at the consumption center, is used to replace the inoperative unit. If no spare is available at consumption center i, then a backorder occurs. The failed unit is sent to the repair center at the instant of failure, and the shipping time is T_i. The repair center repairs the items according to a general distribution function F whose mean is $1/\mu$. There is no queueing or other interaction between items at the repair center, so that the repair center can be thought of as an infinite-server queueing system.

When the item comes out of repair, we must decide whether to keep it (temporarily) in reserve stock at a repair center, or to ship it to consumption center $i, 1 \leqslant i \leqslant m$. The items which are held in reserve at the repair center may be sent at any time to a consumption center.

EXAMPLE 1: Let $m = 3, \lambda_1 = 0.1, T_1 =$ 10 days, $\lambda_2 = 0.02, T_2 =$ 3 days, $\lambda_3 = 0.3, T_3 = 6.67$ days, F be exponential with a mean of 7 days, and $N = 9$. Let each base have 3 items at time 0. A simulation of the system is shown in Table 1.

TABLE 1

Time	Event	Consequence
0.15	Failure at C.C. 3	Inventory level drops to 2. Item will arrive at repair center at 6.82.
2.37	Failure at C.C. 2	Inventory level drops to 2. Item will arrive at repair center at 5.37.
4.76	Failure at C.C. 3	Inventory level drops to 1. Item will arrive at repair center at 11.43.
5.37	Arrival at repair center	
6.14	Failure at C.C. 1	Inventory level drops to 2. Item will arrive at repair center at 16.14.
6.72	Completion of repair at repair center. Repaired item is sent to C.C. 3 where it will arrive at time 13.39.	

Let $s_i(t)$ be inventory level at consumption center i at time t. Our objective is to find a decision rule which minimizes

$$B = \frac{1}{m} \lim_{t \to \infty} \frac{1}{t} E\left[\sum_{i=1}^{m} \int_0^t s_i(w)^- dw\right] \tag{1}$$

where $x^- = \max(-x, 0)$ and E stands for expected value. The average backorder rate is B, and it depends on the initial state as well as the policy used.

Let $x_i(t)$ be the sum of $s_i(t)$ plus the number of items en route to consumption center i. Rather than use the objective function (1), it is preferable to perform a time translation as was done in [2] and let $c_i[x_i(t)]$ be the expected inventory shortage at consumption center i at time $t+T_i$, given that the inventory level at time t is $x_i(t)$. This term is well-defined since any decisions made after time t will influence shortages at consumption center i only after time $t+T_i$. Thus $c_i[x_i(t)] = E[s_i(t+T_i)^-]$, and the objective function in terms of $c_i[x_i(t)]$ is to minimize

$$B = \frac{1}{m} \lim \frac{1}{t} E\left\{\sum_{i=1}^{m}\int_0^{T_i} s_i(w)^- dw + \int_0^{t-T_i} c_i[x_i(w)]dw\right\} \tag{2}$$

$$= \frac{1}{m} \lim_{t\to\infty} \frac{1}{t} E\left\{\sum_{i=1}^{m}\int_0^{t} c_i[x_i(w)]dw\right\}.$$

Since $s_i(t+T_i) = x_i(T)-D_i$, where D_i has a Poisson distribution with a mean of $\lambda_i T_i$, we can calculate $c_i[x_i(t)]$ by

$$c_i[x_i(t)] = \sum_{j=\max[x_i(t),0]}^{\infty} [j-x_i(t)]\, p(j|\lambda_i T_i), \tag{3}$$

where $p(i|\lambda) = e^{-\lambda}\lambda^i/i!$, the probability that a Poisson random variable with mean λ equals i.

We conclude this section with a section with an important observation about the model. Let $U(t)$ be the number of items en route to the repair center and in repair at time t, the symbol U is supposed to suggest "unavailable." This number is independent of all of our past (and of course future) decisions of where to send the repaired item. Define $A(t)$ to be the number of repaired items held at the repair center, plus the items which are en route to the consumption centers, plus the inventory levels at the consumption centers; the symbol A is used to suggest available. Then $A(t)-\sum_{i=1}^{m} x_i(t)=$ number of items held at the repair center. The basic accounting equation is $U(t)+A(t) = N$.

THEOREM 1: Let the initial condition be $U(0) = 0$. (i) For $t \geqslant 0$, $U(t)$ has a Poisson distribution with an increasing mean of $\lambda t P(t)$, where $\lambda=\sum_{i=1}^{m} \lambda_i$, $P(t) = \frac{1}{t}\int_0^t \sum_{i=1}^{m} \frac{\lambda_i}{\lambda}[1-G_i(w)]dw$, and G_i is the distribution function of T_i plus the repair time [*i.e.*, $G_i(t) = F(t-T_i)$]. (*ii*) $\bar{U} \equiv \lim_{t\to\infty} EU(t) = \sum_{i=1}^{m} \lambda_i(T_i + 1/\mu)$, where $1/\mu$ is the mean of the repair time.

PROOF: The result follows with minor modification from a known result with an infinite server queue (see Ross [3,p.18] for example).

COROLLARY: $\bar{A} = \lim_{t\to\infty} E[A(t)]$ exists.

3. THE MAIN RESULT FOR THE DISTRIBUTION PROBLEM

This version of the Transportation Time Look Ahead policy begins with the marginal analysis problem:

$$\min \sum_{i=1}^{m} c_i(x_i) \tag{4}$$

$$\text{subject to } \sum_{i=1}^{m} x_i = \bar{A},\ x_i \text{ an integer.}$$

The c_i are defined by (3), and $\bar{A}$ is defined by the corollary to Theorem 1. (Since $\bar{A} = N - \bar{U}, \bar{X}$ is an integer only if $\bar{U} = \sum_{i=1}^{m} \lambda_i \ (T_i + 1/\mu)$ is an integer, which is not generally the case, of course.)

In this section we will choose simplicity over generality in three specifics. We will assume that $\bar{A}$ is a nonnegative integer. If this is not done we must resort to terminology such as the greatest integer less than $\bar{A}$, which lengthens all the arguments. Second, for some of our results we will assume a convenient initial state for the model. Third, we will let the system get large in a convenient way.

The problem (4) can be solved by a marginal allocation scheme beginning with $x_i = 0$, $i = 1,\ldots,m$, since the c_i are discretely convex. The optimality conditions are that

$$\max[c_i(x_i^*) - c_i(x_i^* + 1)] \leqslant \min[c_i(x_i^* - 1) - c_i(x_i^*)] \tag{5}$$

and $\sum x_i^* = \bar{A}$, where x_i^*, $i = 1, \ldots, m$, are the maximizing values.

Example 2: We keep the same parameters as Example 1 and calculate x_1^*, x_2^*, and x_3^*. As $\bar{U} = 7, \bar{A} = 2$. The optimal allocation is $x_1^* = 1, x_2^* = 0$, and $x_3^* = 1$. It can be verified that $\max_i [c_i(x_i^*) - c_i(x_i^* + 1)] = 0.594$ for $i = 3$ and $\min_i [c_i(x_i^*) - c_i(x_i^* - 1)] = 0.6221$ for $i = 1$.

The Transportation Time Look Ahead policy makes decisions when an item comes out of repair and when there is a failure at a consumption center. When an item comes out of repair, we determine the consumption center, say center k, which maximizes $\{c_i[x_i(t)] - c_i[x_i(t) + 1]\}$ over consumption centers i such that $x_i(t) < x_i^*$, and send the item to consumption center k. If $x_i(t) \geqslant x_i^*$ for all i, then the item is held at the repair center. The term myopic can be applied to this rule since the c_i are the immediate cost functions.

When a demand occurs at consumption center i, we check if both (a) $x_i(t) < x_i^*$ and (b) there is at least one item held at the repair center. We send an item from the repair center to consumption center i if and only if both these conditions hold. Of course, if both conditions hold, then $x_j(t) \geqslant x_j^*$ for all $j \neq i$. One property of the policy is that if $x_i(0) \leqslant x_i^*$ for all i, then $x_i(t) \leqslant x_i^*$ for all i and t.

The first half of our proof of the optimality of the Transportation Time Look Ahead policy is concerned with showing that for all policies and all starting states satisfying $U(0) = 0$, the average backorder rate B is greater than or equal to $\sum_{i=1}^{m} c_i(x_i^*)$. The idea of the proof is that Theorem 1 ensures that regardless of the policy employed the long-run average of items available is A, and when A items are available the minimum value of the cost rate is $\sum_{i=1}^{m} c_i(x_i^*)$.

Let $C[x(t)] = \sum_{i=1}^{m} c_i[x_i(t)]$ and $C(x^*) = \sum_{i=1}^{m} c_i(x_i^*)$.

LEMMA 1: Let Δ be any value between the right-hand side of (5) and its left-hand side. If $A(t) = \bar{A} + n$, then $C[x(t)] \geqslant C(x^*) - n\,\Delta$.

PROOF: Consider the following variation of (4)

$$\min z = \sum_{i=1}^{m} c_i(y_i)$$

$$\text{subject to } \sum_{i=1}^{m} y_i = \bar{A} + n,\ y_i \text{ an integer.}$$

The value of this problem, z^*, will be less than or equal to $C[x(t)]$ when $\sum_{i=1}^{m} x_i(t) \leqslant A(t) = \bar{A} + n$, since the $x_i(t)$ are a feasible solution to the problem. The inequality $z^* \geqslant C(x^*) - n\,\Delta$ follows from the discrete convexity of the c_i. (Q.E.D)

Let $C(t)$ be the expected value of $C[x(t)]$ conditioned only on $U(0)=0$.

LEMMA 2: For any decision rule,

$$C(t) \geqslant C(x^*) + \{\bar{A} - E[A(t)]\}\Delta.$$

PROOF:
$$C(t) = \sum_{n} C[x(t) | A(t) = \bar{A} + n]\, P[A(t) = \bar{A} + n]$$

$$\geqslant \sum_{n} [C(x^*) - \Delta\, n]\, P[A(t) = \bar{A} + n]$$

by Lemma 1

$$= C(x^*) + \Delta\{\bar{A} - E[A(t)]\}. \qquad \text{(Q.E.D)}$$

THEOREM 2: For any decision rule with a starting state satisfying $U(0) = 0$, $B \geqslant \frac{1}{m} C(x^*)$.

PROOF: By the corollary to Theorem 1, given any $\epsilon > 0$ there is a t_ϵ such that for all $t > t_\epsilon, \bar{A} - E[A(t)] \geqslant -\epsilon$. We have

$$mB = \lim_{t \to \infty} \frac{1}{t} E \int_0^t C(w)\, dw$$

$$= \lim_{t \to \infty} \frac{1}{t} \left[\int_0^{t_\epsilon} C(w)\, dw + \int_{t_\epsilon}^{t} C(w)\, dw \right]$$

$$\geqslant \lim_{t \to \infty} \frac{1}{t} \int_{t_\epsilon} (C(x^*) + \bar{A} - E[A(t)]\, \Delta)\, dw$$

$$\geqslant C(x^*) - \epsilon\, \Delta. \qquad \text{(Q.E.D)}$$

The second part of the proof consists of showing that the Transportation Time Look Ahead policy is optimal when the system gets large, which we define in the following way.

We index on the sequence of problems which are getting large by k, where k represents the number of replications of m consumption centers. That is, for problem k there is one repair center and km consumption centers, with k consumption centers described by the param-

eters (λ_1, T_1), k consumption centers described by the parameters (λ_2, T_2), . . . , and k consumption centers described by the parameters (λ_m, T_m). We let $\bar{U}^k, N^k, \bar{A}^k$, and B^k refer to those terms as applied to problem k, so that $\bar{U}^1 = \bar{U}, N^1 = N, \bar{A}^1 = \bar{A}$, and $B^1 = B$ as $k = 1$ represents the original problem. We set $N^k = kN$. The replication characteristics of problem k imply that $\bar{U}^k = k\bar{U}$. Therefore $\bar{A}^k = k(N-\bar{U}) = k\bar{A}$.

By Theorem 2 the value of kmB^K is greater than or equal to

$$\min \sum_{j=1}^{k} \sum_{i=1}^{m} c_{ij}(x_{ij})$$

$$\text{subject to} \sum_{i} \sum_{j} x_{ij} = k\bar{A},\ x_{ij} \text{ an integer,}$$

where $c_{ij}(\cdot) = c_i$ for $j = 1, \ldots, k$. It is not hard to see that the solution is $x_{ij}^* = x_i^*$, satisfying the optimality conditions (5), and consequently that $B^k \geqslant \frac{1}{m} C(x^*)$.

LEMMA 3: Let $Z^k(t) = [U^k(t) - \bar{U}^k]^+$. Then $EZ^k(t) \leqslant k^{1/2}\bar{U}^{1/2}$.

PROOF: Since $EU^k(t) \leqslant \bar{U}^k$ for all $t \geqslant 0, Z^k(t) \leqslant \{U^k(t) - E[U^k(t)]\}^+$ and $Z^k(t)^2 \leqslant \{U^k(t) - E[U^k(t)]\}^2$. We have $[EZ^k(t)]^2 \leqslant E[Z^k(t)^2] \leqslant E\{U^k(t) - E[U^k(t)]\}^2 = Var\ U^k(t) = k\lambda tP(t) \leqslant k\bar{U}$, by Theorem 1. *Q.E.D.*

THEOREM 3: Let B_*^k be the average cost when there are k replications and we are using the Transportation Time Look Ahead policy. Then $B_*^k \leqslant \frac{1}{m}[C(X^*) + k^{-1/2}\bar{U}^{1/2}]$.

PROOF: We assume that the initial state is $x_{ij}(0) = x_i^*, 1 \leqslant i \leqslant m, 1 \leqslant j \leqslant k$, and that $k(N-\bar{A})$ items are held in reserve at the repair center. With this initial state the Transportation Time Look Ahead policy satisfies $x_{ij}(t) = x_i^*$ whenever $A^k(t) \geqslant \bar{A}^k$. This implies that when $A^k(t) \geqslant A^k$ we incur a cost rate of $kC(X^*)$ and $A^k(t) - \bar{A}^k$ items are held at the repair center. When $A^k(t) < \bar{A}^k$ then the cost rate is less than or equal to $kC(x^*) + \bar{A}^k - A^k(t)$, since $c_i(x_i) - c_i(x_i + 1) \leqslant 1$ for all i and x_i. Thus the cost rate, given $A^k(t)$, is less than or equal to $kC(x^*) + Z^k(t)$. By Lemma 3, $C^k(t) \leqslant kC(x^*) + k^{1/2}\bar{U}^{1/2}$ for all t. Therefore $B_*^k \geqslant \frac{1}{m}[C(x^*) + k^{-1/2}\bar{U}^{1/2}]$.

COROLLARY: Let δ be an arbitrary policy and B_{δ^k} the average cost using δ. Then

$$\lim_{k \to \infty} B_*^k - \hat{B}_\delta^k \leqslant 0.$$

PROOF: Combine Theorem 2 and Theorem 3.

The proof of Theorem 3 does not use the fact that when $A(t) < \bar{A}$ we send an item to the consumption center which maximizes $[c_i(x_i) - c_i(x_i + 1)]$. It only requires that we send it to a center which has $x_i(t) < x_i^*$.

The result does require that the proper stock levels be used. Suppose that we use the Transportation Time Look Ahead policy, but set $x_{ij} = \hat{x}_i, 1 \leqslant j \leqslant k$, where $\sum \hat{x}_i = \bar{A}$, and $C(\hat{X}) > C(X^*)$. The average cost of this policy, $\hat{B}^k$, is greater than or equal to $\frac{1}{m} C(\hat{X})$, so that

$$\lim_{k \to \infty} B^k_* - \hat{B}^k < 0.$$

Theorem 3 also shows that the METRIC distribution model is optimal and we outline the argument.

The METRIC distribution model takes the kN inventory items and allocates them to the repair center and the km consumption centers. For d equals 0,1, . . . , kN, METRIC allocates d items to the repair center and $kN-d$ items to the consumption centers. The allocations to consumption center i are based on a Poisson demand of (the paragraph below Eq. (5) of Sherbrooke [4])

$$\lambda_i(T_i + \Delta(d)) \tag{6}$$

where $\Delta(d)$ is the average number of days an item is backlogged at the repair center, where d items are stocked at the depot, and equals [Eqs. (4-5) of Sherbrooke]

$$\sum_{i=d+1}^{\infty} (i-d)p(i|k\bar{U})/k\ \lambda$$

Once the allocation which minimizes the sum of backorders at the consumption center is determined, METRIC operates in time by shipping to the consumption centers in the order demands occur.

As k goes to infinity and d is set equal to $k\bar{U}$, $\Delta(d)$ goes to zero. The remaining $k\bar{A}$ items are allocated to consumption centers in the same way as the Transportation Time Look Ahead policy (Eq. 3 and 4) since the marginal analysis is based on (6) which is the Poisson demand $\lambda_i T_i$. METRIC satisfies the requirement that items are sent to centers such that $x_i(t) > x_i^*$. Therefore Theorem 3 also shows that METRIC is optimal.

ACKNOWLEDGEMENTS

We thank Steve Lippman for making a number of suggestions on the presentation.

REFERENCES

[1] Miller, B., "A Real Time Metric for the Distribution of Serviceable Assets," RM-5687-PR, The RAND Corporation, Santa Monica, California, (1968).

[2] Miller, B., "Dispatching from Depot Repair in a Recoverable Item Inventory System: On the Optimality of a Heuristic Rule," Management Science, **21** 316-325, (1974).

[3] Ross, S., *Applied Probability Models with Optimization Applications* (Holden-Day, San Francisco, 1970).

[4] Sherbrooke, C., "METRIC: A Multi-Echelon Technique for Recoverable Item Control," Operations Research **16**, 122-141 (1968).

[5] Simon, R., "Stationary Properties of a Two-Echelon Inventory Model for Low Demand Items," Operations Research **19** 761-773 (1971).

PREDICTING THE COST PERFORMANCE OF INVENTORY CONTROL SYSTEMS BY RETROSPECTIVE SIMULATION

Alastair MacCormick*

Department of Management Studies
University of Auckland, New Zealand

ABSTRACT

When an inventory-control system is designed to utilize statistical demand information in computing an (s,S) policy, it may be necessary to use performance forecasts to calibrate the relationship between control parameters and performance measures. Using a simulation model to test prediction by retrospective simulation we estimated the bias and variance of forecasts for expected costs and other operating characteristics. The results show the sensitivity of the prediction bias to changes in demand, cost, and lead time parameters and to choices made in system design.

When a new inventory control system is installed, the design process typically involves forecasting the system's future performance. We investigate the statistical accuracy of forecasts made by employing retrospective simulation. We examine the bias and dispersion in these forecasts and how they vary with system settings, such as the length of the demand history used for policy revision and forecasting, and with environmental specifications, such as the underlying demand process, costs, and replenishment lead time. Extensive tests of the retrospective forecasting technique are made with a computer simulation model.

1. PERFORMANCE FORECASTS IN INVENTORY SYSTEM DESIGN

We assume that the analyst, in designing an inventory-control system, has selected a class of decision rules. Specifically, we consider the (s, S) class of reorder point, reorder level rules. The analyst also must choose a routine to calculate the values of s and S, and these computations depend on the specification of the demand process. If the available empirical information about demand is only a historical sample with a limited number of previous observations, usually the analyst must settle for a computational approach that provides an approximately optimal (minimum expected total cost) statistical policy, since optimal statistical policies are available for only a very restricted class of situations (e.g., [9]). We investigate approximately optimal statistical (s, S) policies that employ the normal distribution, as one simplifying assumption, and estimates of the demand mean and variance computed from recent demand history.

In inventory systems an important trade off is between inventory investment and service, that is, out-of-stock performance. In the classic (s, S) model ([1], [2]), this trade off is influenced by the relative values of the unit holding and penalty cost parameters, the latter

*Currently Visiting Professor at Graduate School of Business Administration, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

often being specified judgmentally. In fact, the assessment of the appropriate penalty cost value may be done, partly, in view of its impact on inventory investment. For this and related reasons, the analyst, prior to implementing an inventory-control system, wishes to predict the system's performance for one or more specifications of the parameter settings employed in the computational process.

For an (s, S) type statistical system, the various approximations that are employed in the computations make it highly questionable to use probability formulas to provide accurate values for the desired system performance forecasts of inventory investment, replenishment frequency, service, and corresponding costs. Therefore, the analyst usually resorts to retrospective simulation. That is, the analyst reuses the limited amount of past data, from which the demand mean and variance are estimated and employed in the (s, S) computations, to predict (simulate) how the chosen rules will perform in the future. On the basis of the resultant statistical forecasts, the analyst may adjust the penalty cost or other parameters. The trade off process via retrospective simulation typically is carried out with aggregates for one or more groups of inventoried items.

The common practice of employing retrospective simulation to provide forecasts of future system's performance raises the following questions.

1. How good are the statistical predictions of the system's future operating characteristics? Are the forecasts biased? What is their variability?

2. How does the accuracy of retrospective simulation forecasts depend on the amount of historical demand information available? On the system's parameter specifications? On the demand environment?

For a simplified one-period inventory model, MacCormick [6] showed that forecasts by retrospective simulation do have prediction biases (for example costs are underestimated), and analyzed the sensitivity of the bias to underlying model specifications, such as mean demand, cost parameters, and length of demand history.

Here we treat a multiperiod environment and the complexity of the system necessitates using computer simulation (MacCormick [7], Estey and Kaufman [4]) to investigate the questions posed above. We describe the inventory model in the next section and the design for the simulation experiment in Section 3. Subsequently, in Section 4, by extensive simulation testing, we analyze the properties of the forecasts.

2. A SINGLE-ITEM MODEL

Before recognizing the actual statistical environment that the system's designer faces, we describe an "ideal" single-item model that motivates the approach. Consider the (s, S) model given in Veinott and Wagner [10]: every period the stock level (on hand plus on order) is reviewed. Demand in each period is assumed to be independently, identically distributed, with mean μ and variance σ^2. Stockouts are backlogged and eventually filled, and replenishments arrive a known fixed lead time λ periods after ordering. There is a fixed ordering cost C_{fix} for each replenishment action. (The purchase cost is directly proportional to the quantity ordered. Since the optimality criterion is in terms of expected cost per period and all demand is eventually satisfied, this cost component can be omitted from further consideration.) The charges applied to period-end inventory and backlog are linear, with unit holding cost C_{in} and unit backlog penalty C_{out}, respectively. In any period the time sequence of events is review and order, delivery, and demand. This stationary model assumes an unbounded horizon.

Under the above ideal assumptions, an optimal rule is of the (s, S) form: if, on review, inventory-on-hand-and-on-order x is below s, an order is placed for a replenishment quantity of $S-x$ units.

Wagner [12] gives an algorithm to compute approximately optimal (s, S); it is adapted from Robert's [8] algorithm by assuming that the upper tail of the lead time demand distribution is approximated by the normal distribution and incorporating a heuristic modification proposed by Wagner, O'Hagen and Lundh [13] for situations of frequent replenishment. Let $P(\cdot)$ be the standardized cumulative normal distribution, and $I(\cdot)$ be the standardized normal loss function:

$$P(u) = (2\pi)^{-1/2} \int_{-\infty}^{u} e^{-x^2/2}\, dx,$$

$$I(u) = (2\pi)^{-1/2} \int_{u}^{\infty} (x-u)\, e^{-x^2/2}\, dx.$$

Further let the Wilson lot size be denoted as:

$$Q_w = (2\mu\ C_{\text{fix}}/C_{\text{in}})^{1/2},$$

and let $\mu_\lambda = (\lambda+1)\mu$ and $\sigma_\lambda^2 = (\lambda+1)\sigma^2$ denote the mean and variance of demand over the lead time plus one period (the interval between reviews). Values for u and v are set by the solutions to

$$I(u) = C_{\text{in}} Q_w/(C_{\text{out}}\sigma_\lambda),$$

$$P(v) = C_{\text{out}}/(C_{\text{in}} + C_{\text{out}}).$$

If Q_w/μ is greater than 1.5, the (s, S) pair is determined by $s = \mu_\lambda + u\sigma_\lambda$ and $S = s + Q_w$; otherwise,

$$s = \mu_\lambda + \text{minimum } (u\sigma_\lambda, v\sigma_\lambda),$$

$$S = \mu_\lambda + \text{minimum } (u\sigma_\lambda + Q_w,\ v\sigma_\lambda).$$

With discrete demands, the values for s and S are rounded to the nearest integer.

The above approximately optimal values for s and S utilize only μ and σ^2 in specifying the demand distribution. The system's designer, therefore, can adapt the formulas to a statistical environment by employing estimates of μ and σ^2 based on the available demand history.

Now we elaborate on the design approach in the statistical environment. Suppose that the analyst has available n periods of demand history, and believes that the next n periods of demand will be drawn from the same underlying demand process (at least approximately). (Actually, n may be a design choice variable, and would reflect the analyst's view of the limited relevance of historical data.) The analyst, then, statistically estimates μ and σ^2 from the past n periods and calculates the (s, S) rule for the next n periods. To forecast performance over the next n periods by retrospective simulation, the analyst initializes the system using the current value of inventory-on-hand-and-on-order, and operates the system with the calculated (s, S) rule and the previous n demands.

The system, when implemented, would operate in this fashion over an extended length of time, that is, every n periods, the latest n demands would be utilized to calculate (s, S) for the next n periods.

3. EXPERIMENTAL DESIGN

The forecasting method just described has been tested for a large number of different specifications of the inventory model.

Table 1 exhibits the full factorial design adopted to specify simulation inputs. There are 648 combinations, all of which were tested. If unit time is one week, the demand history length values of 13, 26, or 52 correpond to quarterly, semiannual, or annual intervals.

TABLE 1. *Input specifications: a full factorial design*

Factor	Levels	Number of Levels
Demand distribution Mean = μ Variance = σ^2	Negative binomial (σ^2/μ=9) Negative binomial (σ^2/μ=3) Poisson (σ^2/μ=1)	3
Mean demand (μ)	2,4,8,16	4
Unit holding cost (C_{in})	1	1
Unit backlog penalty cost (C_{out})	4,9,99	3
Replenishment setup cost (C_{fix})	32,64	2
Replenishment lead time (λ)	0,2,4	3
Demand history length (n)	13,26,52	3

For each of the 648 cases, we replicated the retrospective forecasts 200 times to derive the comparisons between actual and predicted values of several operating characteristics, including expected costs and backlog frequency. In each replication, the sample values for the actual and predicted performance characteristics are averages over the n periods of that replication. In estimating actual expected costs and service performance, we made use of the conditional expectation method of variance reduction (Ehrhardt [3]). To provide a set of initial conditions for each replication, we employed ergodic statistical theory in the experiment as follows. For each case, we let the 200 replications comprise a single historical time sequence. That is, the (s,S) policy values were computed at periods $n+1, 2n+1, 3n+1, \ldots$, and the retrospective forecasts were based on the initial inventory position at each of those periods. Thus, the entire set of replications for a case represents $200n$ consecutive time periods. We treated the sequence of 200 measurements on each operating characteristic as a covariance stationary series, and derived estimates for the sampling variance by using an autoregression model (Appendix B of Fishman [5]).

4. FORECASTS BY RETROSPECTIVE SIMULATION

Before presenting summary results for the 648 cases specified in Table 1, we examine detailed results for a single case to get initial insight to the questions posed in section 1. The case is

Negative binomial demand ($\sigma^2/\mu=9$)

Mean demand $\mu = 16$

Unit backlog cost $C_{out} = 99$

Replenishment set up cost $C_{fix} = 64$

Replenishment lead time $\lambda = 4$.

Observe that each parameter is at its largest value in Table 1. For this single special case, we employed 1200 replications, instead of only 200. Figure 1 shows five different operating characteristics. On each plot the horizontal axis represents the demand history length n. Nine different lengths are tested, ranging from 6 to 52. An asterisk (*) designates the estimate of the actual n-period expected value of the operating characteristic; the corresponding cross (x) shows the estimated average forecast value from retrospective simulation. Each estimate is an average from 1200 n-period replications. (Table A1 of the Appendix provides these estimates of the expected value, along with the standard deviation of the n-period average operating characteristic.) The prediction bias is represented by vertical differences between the pairs of curves in each plot of Figure 1.

The horizontal line on each plot shows the level of the expected value of the operating characteristic when control is by the normal approximation policy (Wagner [12]), with the underlying demand mean 16 and variance 9×16 used to calculate the (s,S) parameters. Asymptotically, the actual and forecast characteristics for the statistical version tend to this value as the demand history length increases, assuming that demands from consecutive pairs of n periods are drawn from the same underlying distribution.

For each of the operating characteristics, Figure 1 shows that the actual expected value is systematically underestimated by a retrospective simulation forecast. The bias is negligible for inventory quantity and replenishment frequency, even for n as small as 6 periods. But the bias is severe for the backlog quantity and frequency, even for n as large as 52 periods. Most of the prediction bias for expected total cost per period can be traced to the bias in the forecast of backlog quantity. For each characteristic, the bias becomes smaller with increases in the demand history length. With demand history length set at 13 periods, the value of expected total cost is 179% of the expected value of the forecast. Likewise, the value of expected period-end inventory is 101%, backlog frequency is 1174%, expected backlog is 2848%, and replenishment frequency is 101%, of the respective expected forecast value. With demand history length at 52 periods, the corresponding ratios are: expected total cost 122%, period-end inventory 100%, backlog frequency 197%, expected backlog 273%, and replenishment frequency 100%.

Note that the results in Figure 1 pertain to the expected values of the actual and forecasted n-period-average operating characteristics. It is also of interest to determine how accurate a single forecast may be, by examining the distribution of the forecast error. To discuss accuracy, we distinguish two values as possible objects of the forecast. When making a forecast for the short-run performance of a policy being evaluated for one item, the object is to forecast the exact n-period-average value which each operating characteristic will take when the policy is implemented over the next n periods. In the second case, we envisage an ensemble of an arbitrarily large number of items with identical underlying specifications, no interdependence of

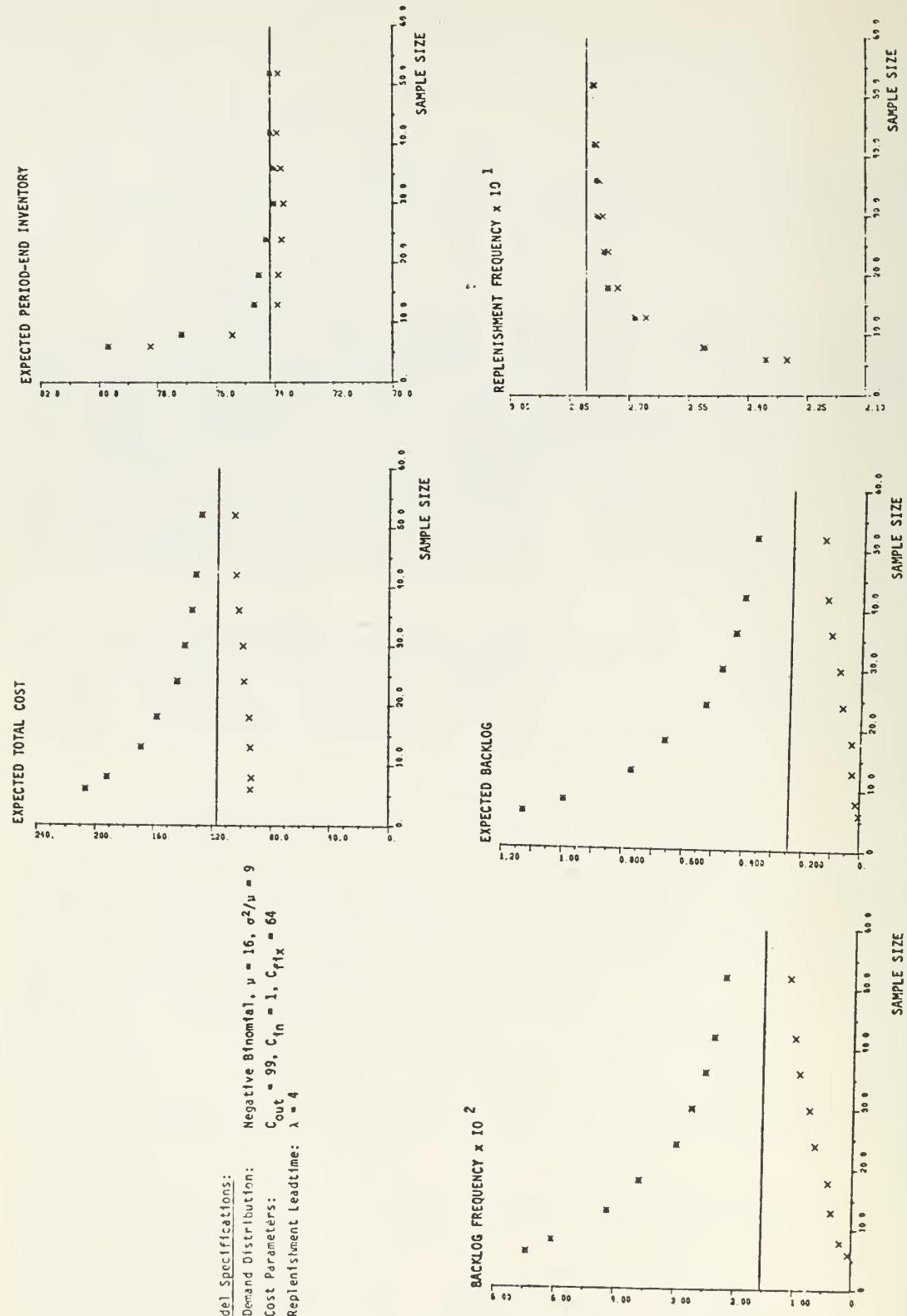

FIGURE 1. Estimates of the expected values of several operating characteristics and their average forecast values, for various demand sample sizes.

demand and initial inventory positions randomly sampled from an equilibrium probability distribution. The object of the forecast is to predict from the n-period history of an arbitrarily chosen item the ensemble average or expected operating characteristics. By ergodic theory, these expected values can be estimated as the long run operating characteristics obtained for one item by time averaging when the policy is revised repeatedly every n periods in a stationary environment. This second case will yield some insight to performance prediction for inventory systems with many items.

To measure the accuracy of a single forecast, for each operating characteristic we constructed from the 1200 replications a sample distribution of the forecast error, expressed as a percentage relative to the forecast. Tables 2 and 3 show these distributions at selected values of the error percentage. Consider first the error distributions for n-period-average total cost. When the object is a short-run forecast for one item and the history length is 13 periods, Table 2 shows an estimated 64.4% chance that the actual 13-period-average total cost will be outside the range 10% below to 10% above the forecast value with odds of 51.2 to 13.2 (approximately 4 to 1), that the forecast will be on the low side of that actual value. When the history length is 52 periods, the chance that the forecast of the 52-period-average total cost will be in error by 10% or more of itself is 62.5%, much the same as for $n=13$, but the odds for an underestimate

TABLE 2. *Percentage error distribution for forecasts of short-run operating characteristics*

Table 2 shows empirical cumulative distributions of the percentage difference, measured relative to the *forecast*, between the forecast and actual n-period-average values realized for each operating characteristic. The distributions are each obtained from 1200 replications for the single item case having specifications: negative binomial demand ($\mu = 16$, $\sigma^2/\mu = 9$); $C_{out} = 99$, $C_{fix} = 64$, $C_{in} = 1$ and $\lambda = 4$.

Operating Characteristic	Demand History Length (n)	Percentage of Forecast by which Actual Value is Underestimated is at Least:					Forecast Exactly On Target	Percentage of Forecast by which Actual Value is Overestimated is at Least:				
		100%	50%	20%	10%	0		0	10%	20%	50%	100%
Average Total Cost	13	14.8	19.2	34.9	51.2	70.3		29.7	13.2	05.1	0.8	
	52	12.3	19.4	29.2	41.7	63.7		36.3	20.8	12.6	1.5	
Average Period-End Inventory	13		3.6	25.3	39.3	51.8		48.2	36.5	26.3	4.9	
	52			07.9	24.1	51.6		48.4	25.7	09.3	0.3	
Average Backlog Quantity	13	22.3[1]	22.4	22.5	22.5	22.5	75.2	02.3	02.3	02.3	02.2	02.2[2]
	52	36.7[1]	37.6	38.0	38.2	38.6	32.7	28.7	28.7	28.4	27.1	24.3[2]
Backlog Frequency	13	21.9[1]	21.9	21.9	21.9	21.9	75.9	02.2	02.2	02.2	02.1	2.1[2]
	52	35.7[1]	36.2	36.3	36.3	36.3	38.0	25.7	25.7	25.7	25.6	23.5[2]
Replenishment Frequency	13	4.5	12.4	31.7	31.7	31.7	36.3	32.0	32.0	32.0	03.4	
	52		0.04	10.1	21.1	39.7	19.8	40.5	23.2	04.9		

1) (See Table 4). A zero forecast was followed by a positive n-period average backlog in 253 (21.1%) of the 1200 replications for n = 13, and in 332 (27.7%) of the replications for n = 52.

2) (See Table 4). A backlog was predicted but did not eventuate in 25 (2.1%) of the 1200 replications for n = 13, and in 282 (23.5%) of the replications for n = 52.

TABLE 3. *Percentage error distribution for forecasts of expected operating characteristics*

Table 3 shows empirical cumulative distributions of the percentage difference, measured relative to the *forecast*, between the forecast and the *expected* value of each operating characteristic. The distributions are each obtained from 1200 replications for the single item case having negative binomial demand ($\mu = 16$, $\sigma^2/\mu = 9$); $C_{out} = 99$, $C_{fix} = 64$, $C_{in} = 1$ and $\lambda = 4$.

Operating Characteristic	Demand History Length (n)	Forecast Underestimates by at Least:					Forecast Overestimates by at Least:				
		100%	50%	20%	10%	0	0	10%	20%	50%	100%
Total Cost	13	39.6	81.4	96.8	98.2	99.0	01.0	00.6	00.6	0.1	
	52	00.2	23.7	72.7	82.4	88.2	11.8	08.0	04.7	0.3	
Period-End Inventory	13	00.8	07.9	30.3	42.8	57.2	42.7	27.7	15.1	0.2	
	52			09.3	27.7	55.3	44.7	19.5	03.8		
Backlog Quantity	13	97.6*	98.2	98.4	98.7	98.7	01.3	00.8	00.7	0.6	
	52	79.0*	82.8	86.5	87.3	88.9	11.1	09.9	08.4	3.6	
Backlog Frequency	13	96.3*	96.3	96.3	96.3	96.3	03.7	03.7	03.7	3.7	
	52	59.8*	59.8	59.8	59.8	87.0	13.0	13.0	13.0	3.2	
Replenishment Frequency	13	01.8	12.0	12.8	49.6	49.6	50.4	50.4	08.8		
	52			02.8	18.7	50.4	49.6	03.5			

*(See Table 4) Zero backlog was predicted in 1155 (96.3%) of the 1200 replications for n = 13, and in 718 (59.8%) of the replications for n = 52.

are only about 2 to 1. Table 3 demonstrates in similar fashion the errors when the object is to forecast the ensemble average total cost. With the 13-period demand history the chance the forecast is in error by 10% of itself, or more, is estimated to be 98.8%, an underestimate being almost certain. With as many as 52 demands in the history, the chance is still high at 90.4%, and the odds for underestimation are about 10 to 1.

Turning to operating characteristics which form the components of total cost, there is substantial dispersion in the forecasts for the average period-end inventory and the replenishment frequency, but it is the prediction error in the backlog quantity, weighted by the unit backlog penalty of 99, which is the main source for the bias and dispersion observed in the forecasts of average total cost. Even with 52 periods of history, the chances of a short-run forecast (Table 2) being 10% of itself or more in error are 49.8% for inventory, 66.9% for backlog and 44.3% for the replenishment frequency, and when forecasting expected values (Table 3), the equivalent chances are 47.2%, 97.2% and 22.2%.

We found wide dispersion in percentage forecast error to be very common in the 648 cases of Table 1. Indeed the dispersion is so great that little reliance can be given to a single forecast made from the demand history for one item. This conclusion applies with particular force to forecasts of backlog performance, which we next examine in greater detail, before considering whether an aggregation of single-item predictions can be useful for predicting systems performance.

A reliable calibration of the important trade off between inventory investment and backlog performance can be made only if there is an accurate method to forecast backlogs. But inventory control systems are designed to make backlogs relatively rare events: indeed for the special case we have been using for illustration, when the demand history length is 13 periods, on the average a backlog actually occurs about once in 24 periods; with a history length of 52 a backlog happens about once in 43 periods. Using demand samples with history lengths so short relative to the average times between backlogs, it is hardly surprising that it is very difficult to predict correctly the backlog frequency, let alone the magnitude of any backlogs. Table 4 illustrates this problem for backlog frequency. As an example from Table 4(a), which shows results for 1200 replications with the demand history length set at 13 periods, consider the 126 in which a backlog actually occurred once in 13 periods. In only 9 of these 126 replications was the one backlog predicted correctly. Of the remaining forecasts, 116 predicted no backlog and one predicted two backlogs. The nonzero off-diagonal entries in Table 4 represent replications in which the forecast did not correctly predict the actual frequency, and show that when backlogs occurred there was little chance of an accurate prediction. Indeed, for the case $n=52$, a chi-square test on the sample distribution (aggregated to 5 rows and 4 columns) fails to reject the hypothesis that forecast and actual frequencies are statistically independent, at a level of significance of 90%.

TABLE 4. *Sample frequency distributions for actual and forecast backlog frequency*

I = number of periods with backlog during n-period retrospective simulation.
J = number of periods with backlog during subsequent n periods.

(a) HISTORY LENGTH n=13

FORECAST

ACTUAL	I=0	1	2	3	Total
J=0	902	19	4	2	927
1	116	9	1		126
2	48	2			50
3	40	2			42
4	19	1			20
5	16	1	1		18
6	7	1			8
7	1				1
8	5	2			7
9					
10					
11	1				1
Total	1155	37	6	2	1200

(b) HISTORY LENGTH n=52

FORECAST

ACTUAL	I=0	1	2	3	4	Total
J=0	386	189	76	13	4	668
1	120	51	17	4	2	194
2	83	29	12	1	2	127
3	50	27	6	6		89
4	31	14	1	1	1	48
5	16	4	2	1		23
6	10	5	2			17
7	6	3		1		10
8	8	1	2			11
9	1	1		1		3
10	2	1		1		4
11	1	1				2
12	3					3
13						
14	1					1
Total	718	326	118	29	9	1200

We consider now whether single item forecasts can be made more useful for systems predictions by aggregation over items. Typically, a representative sample of items from the system is employed to make aggregate forecasts. As more items are included in this sample, provided there is no interdependence between items and all have the same sign for forecast bias, the bias of the aggregate forecast increases at a rate faster than the standard deviation of the aggregate. Given a sufficiently large number of items, correction for bias will reduce appreciably the mean square error of the system forecast. Since the bias may be sensitive to the parameters (cost, demand, lead time) for each item in the sample, and will impact the aggregate accordingly, we next explore the sensitivity of the bias in forecasts aggregated for different groups of items.

We examine 9 groups of 72 items formed by partitioning the 648 cases in Table 1, so that each item in a group has the same demand distribution, variance-to-mean ratio, and history length. In Table 5, we show the extent to which prediction bias is sensitive to the demand history length and the demand distribution. To form the 72 items aggregates, we have used cost components, which facilitate sensitivity comparisons. As we saw in the single-item case above, the bias decreases as the history length increases. By comparing rows in the table, we see that

TABLE 5. *Bias in Forecasts of Aggregate Operating Characteristics by Retrospective Simulation*

Each actual and forecast value is a sample average from 200 replications, obtained by partitioning the 648 items specified in Table 1 into 9 groups, each with a given demand distribution and demand history length and by aggregating within each group.

Demand History Length n:	n = 13			n = 26			n = 52		
	Actual Value	Forecast Value	Bias (%)	Actual Value	Forecast Value	Bias (%)	Actual Value	Forecast Value	Bias (%)
Negative Binomial Demands $\sigma^2/\mu = 9$									
Holding Cost	1874	1822	(02.8)	1817	1788	(01.6)	1804	1790	(00.8)
Backlog Cost	1562	280	(82.0)	1279	458	(64.2)	1080	598	(44.6)
Replenishment Cost	670	658	(01.8)	682	674	(01.2)	683	679	(00.6)
Total Cost	4105	2760	(32.8)	3778	2920	(22.7)	3567	3066	(14.0)
Backlog Frequency	.108	.070	(34.6)	.100	.080	(20.1)	.093	.083	(11.5)
Negative Binomial Demands $\sigma^2/\mu = 3$									
Holding Cost	1290	1271	(01.5)	1274	1262	(00.9)	1270	1265	(00.5)
Backlog Cost	564	186	(66.9)	457	242	(47.1)	401	289	(28.0)
Replenishment Cost	744	737	(01.1)	750	745	(00.6)	750	747	(00.3)
Total Cost	2598	2194	(15.6)	2480	2249	(09.3)	2421	2301	(05.0)
Backlog Frequency	.101	.082	(23.7)	.094	.083	(13.2)	.092	.086	(06.9)
Poisson Demands $\sigma^2/\mu = 1$									
Holding Cost	1024	1019	(00.6)	1020	1018	(00.3)	1019	1017	(00.2)
Backlog Cost	245	133	(45.9)	207	149	(27.9)	190	162	(14.6)
Replenishment Cost	775	771	(00.5)	777	776	(00.2)	777	776	(00.1)
Total Cost	2046	1922	(06.0)	2005	1943	(03.1)	1986	1956	(01.5)
Backlog Frequency	.093	.082	(13.0)	.090	.084	(06.5)	.088	.085	(03.0)

the bias decreases as the group demand variance-to-mean ratio decreases. Most of the bias in total cost is attributable to the backlog component.

Tables A2, A3, and A4 of the Appendix further split each group of 72 items to show the sensitivity of the prediction biases to other system parameters. The bias in the forecast of expected average total cost increases with increases in C_{out}/C_{in}, replenishment lead time λ and demand μ, but decreases with C_{fix}/C_{in}. The prediction bias is large for the backlog component of expected total cost, but negligible for the holding and replenishment components. More information about the properties of aggregate forecasts is available in MacCormick [6] and Estey and Kaufman [4].

We conclude that forecasts made by retrospective simulation have prediction biases, leading to underestimates of operating-characteristic expected values. The underestimation is particularly significant for backlog performance measures. The bias is due to the double use of historical demand information in the statistical approximation policy: first to set the system (s,S) parameters, and then to forecast the subsequent performance. Further, there is considerable dispersion in the distribution of forecast errors for individual predictions. Hence, the system's designer should be most cautious in making system's predictions, especially with regard to service performance and the trade off between inventory investment and service. The results in this paper suggest that predictions bases on only a small sample of items may have a large error. Thus, the analyst is advised to be generous in selecting the number of items to use in a system's design test. Similarly, the analyst is advised to use as long a demand history length as is available and sufficiently representative of the future demand environment.

ACKNOWLEDGMENTS

This research was sponsored by the Office of Naval Research (N00014-75-C-0241), the U.S. Army Research Office (DAHC04-75-G-0079), Yale University, and the Universities of Auckland, New Zealand, and North Carolina at Chapel Hill. The author gratefully acknowledges the stimulating contributions of Harvey M. Wagner and Matthew J. Sobel, and computational help from Arthur S. Estey and Ronald L. Kaufman.

REFERENCES

[1] Arrow, K., T.M. Harris, and J. Marschak, "Optimal Inventory Policy," Econometrica 19, 250-272 (1951).

[2] Arrow, K., S. Karlin, and H.E. Scarf, *Studies in the Mathematical Theory of Inventory and Production*, (Stanford University Press, Stanford, California, 1958).

[3] Ehrhardt, R., "Variance Reduction Techniques for an Inventory Simulation," ONR and ARO Technical Report 4, School of Organization and Management, Yale University, New Haven, Connecticut (September 1975) 24 pp.

[4] Estey, A.S., and R.L. Kaufman, "Multi Item Inventory System Policies Using Statistical Estimates: Negative Binomial Demands (Variance/Mean = 9)," ONR and ARO Technical Report 3, School of Organization and Management, Yale University, New Haven, Connecticut (September 1975), 85 pp.

[5] Fishman, G.S., *Concepts and Methods in Discrete Event Digital Simulation*, (Wiley, New York, 1973).

[6] MacCormick, A., "Statistical Problems in Inventory Control," ONR and ARO Technical Report 2, School of Organization and Management, Yale University, New Haven, Connecticut (December 1974) 244 pp.

[7] MacCormick, A., "The One-Period Inventory Model: Control with Imperfect Information," New Zealand Journal of Operational Research *2*, 21-36 (1976).
[8] Roberts, D., "Approximations to Optimal Policies in a Dynamic Inventory Model," Chapter 13 in *Studies in Applied Probability and Management Science*, K. Arrow, S. Karlin and H.E. Scarf, eds. (Stanford University Press, Stanford, California, 1962).
[9] Scarf, H.E., "Bayes Solutions of the Statistical Inventory Problem," Annals of Mathematical Statistics *30*, 490-508 (1959).
[10] Veinott, A.F., and H.M. Wagner, "Computing Optimal (s,S) Inventory Policies," Management Science *11*, 525-552 (1965).
[11] Wagner, H.M., *Statistical Management of Inventory Systems* (Wiley, New York, 1962).
[12] Wagner, H.M., *Principles of Operations Research*, 2nd ed. (Prentice-Hall, Englewood Cliffs, N.J., 1975).
[13] Wagner, H.M., M. O'Hagen, and B. Lundh, "An Empirical Study of Exact and Approximately Optimal Inventory Policies," Management Science *11*, 690-723 (1965).

APPENDIX

ADDITIONAL OPERATING CHARACTERISTIC DATA

Table A1 contains the detailed operating characteristic data obtained for the single item example discussed in Section 4. Tables A2, A3, and A4 display additional sensitivity data for bias in cost forecasts.

TABLE A1. *Dependence of actual and forecast operating characteristics on size of demand sample used at revision*

Demand Sample Size n	Estimated Means and Standard Deviations, Aggregated over n Periods, of:									
	Total Cost		Period-End Inventory		Period-End Backlog		Backlog Frequency		Replenishment Frequency	
	Actual	Forecast	Actual	Forecast	Actual	Forecast	Actual	Forecast	Actual	Forecast
6	206.4 (456.4)	93.1 (31.7)	79.7 (44.5)	78.2 (29.5)	1.127 (4.760)	0.002 (0.040)	0.0546 (0.1559)	0.0006 (0.0096)	0.236 (0.139)	0.230 (0.124)
8	191.5 (336.4)	92.9 (29.8)	77.2 (40.5)	75.4 (25.7)	0.993 (3.530)	0.014 (0.148)	0.0503 (0.1323)	0.0020 (0.0164)	0.251 (0.114)	0.251 (0.099)
13	168.0 (238.9)	93.6 (26.5)	74.7 (32.5)	73.9 (19.1)	0.769 (2.526)	0.027 (0.196)	0.0411 (0.1002)	0.0035 (0.0194)	0.268 (0.073)	0.266 (0.067)
18	157.3 (165.4)	94.3 (23.4)	74.5 (28.8)	73.9 (17.1)	0.658 (1.771)	0.030 (0.155)	0.0359 (0.0747)	0.0041 (0.0158)	0.275 (0.056)	0.273 (0.049)
24	143.7 (124.1)	97.6 (29.7)	74.3 (24.6)	73.8 (14.2)	0.523 (1.342)	0.063 (0.259)	0.0297 (0.0628)	0.0064 (0.0181)	0.276 (0.046)	0.275 (0.040)
30	138.4 (110.2)	98.8 (28.0)	74.0 (22.5)	73.7 (12.8)	0.470 (1.198)	0.074 (0.244)	0.0272 (0.0543)	0.0074 (0.0175)	0.278 (0.039)	0.277 (0.032)
36	134.0 (100.6)	101.8 (32.1)	74.0 (20.2)	73.8 (11.4)	0.426 (1.094)	0.104 (0.290)	0.0250 (0.0466)	0.0091 (0.0176)	0.278 (0.035)	0.277 (0.028)
42	131.4 (92.2)	103.6 (31.9)	74.1 (18.9)	73.9 (10.8)	0.398 (1.005)	0.120 (0.290)	0.0235 (0.0421)	0.0100 (0.0175)	0.279 (0.032)	0.278 (0.026)
52	127.8 (79.5)	104.8 (29.2)	74.2 (17.3)	73.9 (9.6)	0.361 (0.874)	0.132 (0.266)	0.0217 (0.0372)	0.0110 (0.0158)	0.279 (0.029)	0.278 (0.022)

Model Specifications:

Demand distribution: Negative binomial with $\mu = 16$, $\sigma^2/\mu = 9$;
Cost parameters: $C_{out} = 99$, $C_{fix} = 64$, $C_{in} = 1$;
Replenishment lead time: $\lambda = 4$.

TABLE A2. *Percentages by which forecast underestimates expected costs per period, for various classifications of 72 items under statistical control (Demand History Length: 13 Periods)*

Cost Component	Total	Input Parameters											
		C_{out}/C_{in}			C_{fix}/C_{in}		Lead time λ			Mean Demand μ			
		4	9	99	32	64	0	2	4	2	4	8	16
Negative Binomial Demands $\sigma^2/\mu = 9$													
Holding Cost	2.8	5.8	3.6	1.0	3.2	2.4	0.2	2.3	4.7	1.7	3.0	3.0	3.0
Backlog Cost	82.0	45.3	65.7	94.0	82.7	81.4	58.9	83.2	89.8	87.8	84.6	80.3	78.3
Replenishment Cost	1.8	1.5	1.9	2.0	1.4	2.1	2.8	1.7	0.9	4.2	2.7	1.3	0.9
Total Cost	32.8	14.9	19.6	46.1	34.3	31.4	17.8	33.5	40.3	40.5	37.1	31.6	27.8
Negative Binomial Demands $\sigma^2/\mu = 3$													
Holding Cost	1.5	3.5	2.1	0.2	1.8	1.3	0.2	1.3	2.6	1.8	1.7	1.4	1.4
Backlog Cost	66.9	30.7	53.0	69.5	69.3	64.5	37.7	65.4	77.8	75.1	73.8	66.7	57.7
Replenishment Cost	1.1	1.3	1.0	0.9	0.9	1.1	1.3	1.0	0.9	2.2	1.0	1.1	0.6
Total Cost	15.6	8.0	10.7	23.1	17.8	13.7	5.7	14.9	22.7	20.8	18.9	15.8	11.5
Poisson Demands $\sigma^2/\mu = 1$													
Holding Cost	0.6	1.6	0.6	0.0	0.4	0.7	0.2	0.2	1.2	0.8	0.8	0.5	0.4
Backlog Cost	45.9	18.0	36.1	81.7	50.6	41.4	16.6	44.0	58.2	53.5	49.5	44.5	42.2
Replenishment Cost	0.5	0.3	0.7	0.6	0.7	0.4	0.7	0.4	0.4	1.4	1.0	0.5	0.0
Total Cost	6.0	3.5	4.4	9.0	7.3	5.0	1.7	5.5	10.0	7.5	7.0	5.9	5.1

TABLE A3. *Percentages by which forecast underestimates expected costs per period, for various classifications of 72 items under statistical control (Demand History Length: 26 Periods)*

Cost Component	Total	Input Parameters											
		C_{out}/C_{in}			C_{fix}/C_{in}		Lead time λ			Mean Demand μ			
		4	9	99	32	64	0	2	4	2	4	8	16
Negative Binomial Demands $\sigma^2/\mu = 9$													
Holding Cost	1.6	3.1	2.1	0.6	1.9	1.3	0.0	1.5	2.6	1.7	1.8	1.5	1.5
Backlog Cost	64.2	28.0	44.9	79.0	65.7	62.7	37.1	63.0	76.3	74.4	68.3	59.8	57.4
Replenishment Cost	1.2	1.0	1.4	1.3	0.9	1.5	1.3	1.5	0.9	4.2	1.4	0.8	0.4
Total Cost	22.7	8.7	12.4	34.0	24.8	20.9	10.3	22.2	30.3	34.5	27.0	20.3	16.8
Negative Binomial Demands $\sigma^2/\mu = 3$													
Holding Cost	0.9	2.1	1.0	0.4	1.0	0.9	0.2	0.9	1.5	1.3	0.9	0.9	0.8
Backlog Cost	47.1	17.3	32.5	71.5	50.1	44.1	19.3	45.7	59.6	58.3	46.0	44.9	43.2
Replenishment Cost	0.6	0.6	0.7	0.5	0.5	0.6	0.6	0.5	0.7	1.9	0.6	0.2	0.3
Total Cost	9.3	4.4	5.8	14.8	10.9	8.1	2.8	9.3	14.1	14.9	9.3	8.6	7.7
Poisson Demands $\sigma^2/\mu = 1$													
Holding Cost	0.3	0.9	0.2	−0.1	0.3	0.2	0.1	0.3	0.3	0.4	0.4	0.3	0.1
Backlog Cost	27.9	7.9	21.4	62.4	31.2	25.1	5.7	27.3	38.5	34.4	27.6	27.2	26.3
Replenishment Cost	0.2	0.2	0.2	0.1	0.1	0.2	0.3	0.2	0.0	0.7	0.4	0.0	−0.1
Total Cost	3.1	1.6	2.3	4.8	3.7	2.6	0.6	3.1	5.2	4.1	3.3	3.0	2.7

TABLE A4. *Percentages by which forecast underestimates expected costs per period, for various classifications of 72 items under statistical control (Demand History Length: 52 Periods)*

Cost Component	Total	Input Parameters											
		C_{out}/C_{in}			C_{fix}/C_{in}		Lead time λ			Mean Demand μ			
		4	9	99	32	64	0	2	4	2	4	8	16
Negative Binomial Demands $\sigma^2/\mu = 9$													
Holding Cost	0.8	1.6	1.0	0.3	0.9	0.7	0.2	0.7	1.3	0.8	0.8	0.9	0.7
Backlog Cost	44.6	16.2	27.8	58.6	45.4	43.9	24.2	43.0	55.8	58.1	49.0	39.2	35.2
Replenishment Cost	0.6	0.8	0.3	0.6	0.5	0.6	0.6	0.5	0.7	1.6	0.7	0.5	0.2
Total Cost	14.0	4.9	6.9	22.1	15.3	12.9	6.5	13.6	19.1	24.9	17.5	11.8	8.8
Negative Binomial Demands $\sigma^2/\mu = 3$													
Holding Cost	0.5	1.2	0.6	0.0	0.5	0.4	0.1	0.4	0.7	0.4	0.9	0.4	0.3
Backlog Cost	28.0	8.4	17.6	47.8	29.3	26.7	10.5	26.6	37.4	36.2	29.3	27.1	23.5
Replenishment Cost	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.3	1.0	0.5	0.2	0.0
Total Cost	5.0	2.2	3.0	8.1	5.7	4.4	1.5	4.8	7.8	8.0	5.8	4.7	3.6
Poisson Demands $\sigma^2/\mu = 1$													
Holding Cost	0.2	0.5	0.3	−0.1	0.2	0.1	0.0	0.1	0.5	0.2	0.4	0.1	0.1
Backlog Cost	14.6	3.7	9.8	39.0	16.6	12.8	1.7	14.1	21.7	19.4	14.2	15.9	11.9
Replenishment Cost	0.1	0.1	0.2	0.0	0.0	0.2	0.1	0.1	0.1	0.5	0.2	0.0	0.0
Total Cost	1.5	0.8	1.1	2.4	1.8	1.3	0.2	1.4	2.8	2.2	1.7	1.6	1.2

ESTIMATING VALUE IN A UNIFORM AUCTION

Francisco J. Samaniego and Lee D. Kaiser*

University of California, Davis
Davis, California

ABSTRACT

Consider an auction in which increasing bids are made in sequence on an object whose value θ is known to each bidder. Suppose n bids are received, and the distribution of each bid is conditionally uniform. More specifically, suppose the first bid X_1 is uniformly distributed on $[0, \theta]$, and the i^{th} bid is uniformly distributed on $[X_{i-1}, \theta]$ for $i = 2, \ldots, n$. A scenario in which this auction model is appropriate is described. We assume that the value θ is unknown to the statistician and must be esimated from the sample $X_1, X_2, \ldots, X_n$. The best linear unbiased estimate of θ is derived. The invariance of the estimation problem under scale transformations is noted, and the best invariant estimate of θ under loss $L(\theta, a) = [(a/\theta) - 1]^2$ is derived. It is shown that this best invariant estimate has uniformly smaller mean-squared error than the best linear unbiased estimate, and the ratio of the mean-squared errors is estimated from simulation experiments. A Bayesian formulation of the estimation problem is also considered, and a class of Bayes estimates is explicitly derived.

1. INTRODUCTION

The bibliography on competitive bidding compiled by R. M. Stark [12] documents the fact that there has been considerable recent interest in mathematical models for bidding processes. For the most part, models proposed for such processes have been used for developing "optimal" bidding strategies. For example, Griesmer and Shubik [4, 5, 6] and Griesmer, Shubik, and Levitan [7] formulate bidding processes as n-person games, and obtain equilibrium solutions under a variety of constraints on the amount of information on previous bids available to each bidder. The only distinction among bidding processes made in our study is between sealed-bid and open-bid auctions. In a sealed-bid auction, each participant submits a bid without knowing the bids of his opponents, while in an open-bid or progressive auction, bids are made sequentially and all participants are fully aware of all previous bids. Models for open-bid auctions are considered in [6]. There seems to be little work on probability models for an observed sequence of bids, and, to our knowledge, inference questions concerning open-bid auctions are essentially untouched. We present in this paper a probability model for an open-bid auction, and we develop and compare several estimators of the parameter of the model.

We will consider an open-bid auction model to be any probability model which generates a sequence of nonnegative, increasing observations $\{X_i\}$. Thus, the joint distribution of the order statistics from any continuous distribution with support in IR^+ could be viewed as an auction model. Similarly, the random-record models studied by D. Graver [3] and by M. Yang [15] could serve as auction models. It is clear that a definitive treatment of modeling and inference

*The second author is now at Colorado State University.

questions for open-bid auctions must deal with the stopping rule in the bidding process that determines the size of the sample of bids to be examined. Much can be learned, however, by examining the estimation problem as a fixed-sample-size problem, and this viewpoint is pursued in section 3. An application is described in which such a simplification seems justified. In section 4, we present a Bayesian formulation of the estimation problem in which the optimal estimator is in fact independent of the stopping rule.

2. THE UNIFORM AUCTION

We will refer to the following probability model as the uniform auction. Let θ be a nonnegative, real-valued parameter to be thought of as the value of the object being auctioned. Let $X_1 \sim U[0, \theta]$, that is, X_1 is uniformly distributed on the interval $[0, \theta]$. For $i = 2, \ldots, n$, let $X_i \sim U[X_{i-1}, \theta]$. Taking $x_0 \equiv 0$, the probability density of the vector of bids $\mathbf{X}$ is given by

$$f(x_1, x_2, \ldots, x_n|\theta) = \begin{cases} \prod_{i=0}^{n-1} \frac{1}{\theta - x_i} & \text{for } 0 < x_1 < \ldots < x_n < \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

This probability model seems to be new in statistical literature, and has a number of interesting characteristics. It is not an order-statistic model, and it differs from the distribution of the order statistics from $U[0, \theta]$ in that the uniform auction typically produces a set of bids tending much more to the right in the interval $[0, \theta]$ than the set of bids from the uniform order-statistic model. The uniform auction has an interesting statistical property that renders the estimation of θ a somewhat delicate problem. One can easily verify by the Lehmann-Scheffe technique [8] that the entire sample of bids $(X_1, \ldots, X_n)$ is a minimal sufficient statistic for θ. Thus, in contrast with many bounded or semibounded order-statistic models, reduction by sufficiency does not accomplish a simplification of the problem.

The probability model considered in this paper is supported by a number of intuitive considerations. As discussed in section 3.3, the value θ in the uniform auction is a scale parameter of the distribution of bids. We are thus able to develop reasonable estimators of θ that are invariant with respect to scale transformations on the data. Such estimators have the property that the estimated values are equivalent regardless of the monetary units in which θ and the bids are expressed. Stated informally, our estimate of θ should not depend on whether the bids are in dollars or in pennies, and each of the estimators obtained in section 3 satisfies our intuition in this regard. It is possible to arrive at the uniform auction rather than other scale-parameter models from an entropy argument, since the uniform distribution on $[X_i, \theta]$ maximizes the entropy among all distributions on the interval. Thus, the uniform auction is a sequential formulation of maximum uncertainty concerning the bidding process.

Precedents exist in the literature on bidding processes for the probability model considered in this paper, as well as for the estimation problem we investigate. In a celebrated paper, Vickrey [14] studied game-theoretic aspects of competitive bidding in a framework in which bids were modeled as uniform random variables. He postulated that, in a sealed-bid auction, a random participant assesses the value of the object at auction according to a uniform distribution on some interval, taken as (0, 1) for simplicity. Under a linear utility function, the unique equilibrium strategy is for each bidder to bid the amount

$$b_i = \frac{N-1}{N} \mathrm{v}_i, \quad i = 1, \ldots, N,$$

where v_i is the value the i^{th} bidder obtains from the uniform distribution and N is the number of bidders. Thus, when bidders employ the equilibrium strategy (which is also Pareto optimal),

each bid is itself a uniformly distributed random variable. In an appendix, Vickrey briefly discusses one possible extension to open-bid auctions — the uniform order-statistic model. We believe that such an extension is inappropriate, since the distribution is not derived using a sequence of conditional distributions based on past bids and is thus not patterned after a real open-bid auction. It is easy to see, however, that the uniform auction is a natural extension to open-bid auctions of Vickrey's sealed-bid model. Suppose each bidder has an initial value assessment v_i, drawn independently from a uniform distribution according to Vickrey's model. If a bidder emerges at random from the N participants, and he submits a bid according to the aforementioned equilibrium strategy, the bid is uniformly distributed. After the first and each successive bid, the bidders reassess the value of the object and, following Vickrey's model, obtain new values according to a uniform distribution on an interval with a different lower bound. Under such a mechanism for reappraisal, the resulting probability model is precisely the uniform auction. The interpretation of θ as the value of the object is consistent with Vickrey's scheme, since it would represent the least upper bound on possible value assessments, and, in theory, one could find a buyer willing to pay more than $\theta - \epsilon$ for the object for any $\epsilon > 0$. An estimation problem in competitive bidding was examined by Christenson [1]. In that paper, a multiparameter probability model was advanced as a distribution of an opponent's bids in a multistage sealed-bid auction. Among the parameters of the model was the value of the object to the bidder. Bayesian linear regression analysis was used for estimating the parameters of the model.

While the uniform auction can be viewed as a natural extension of Vickrey's scheme for value assessment, it can also be motivated from a quite different scenario. Consider an auction with a large number of expert bidders. The annual bull auction at Red Bluff, California, would be an example (see Sosnick [11, page 1310] for a description). Each bidder knows θ, the value (perhaps the market value or perhaps the break-even point in a marketing process after a purchase) of the object at auction. Such an assumption is of course approximately true, at best, but it should be sufficiently close to reality in some auctions to justify the use of our model. In his summary section, Vickrey [14, page 28] discusses the validity of his model, stating

> "... these conclusions are based on a model in which a high degree of rationality and sophistication is imputed to the bidders; nevertheless, in many markets the frequency of the dealings and the professional characteristics of the dealers are such as to make such an assumption not too far from reality."

Given that i bids have been made, the $(i+1)^{st}$ bid is assumed uniformly distributed. This assumption might be justified on behavioral grounds, since the bid that an aggressive bidder would consider would be high in $[X_i, \theta]$, while the bid that a conservative bidder would consider would be low in $[X_i, \theta]$, and hence the random bid that emerges from a large group of bidders might quite reasonably be modeled as a uniform variable. Finally, while the sample size in an open-bid auction tends to provide information about θ, one might consider estimating θ after stopping artifically, that is, after n bids are made, even though the bidding actually continues. Estimating θ from an artifically stopped auction could appropriately be considered a fixed-sample-size problem when n is not too large relative to the expected number of bids. In fact, the primary application of the estimation procedures discussed in the next two sections might well be the estimation of value by a nonprofessional (referred to henceforth as the statistician) based on the early bids of expert bidders.

We now turn to the moment structure of the uniform auction that comes into play in estimating θ. By the recursive relations

$$EX_i = E[E(X_i|X_{i-1})] = \frac{\theta + EX_{i-1}}{2}, \quad i = 2, 3, \ldots, n,$$

together with $EX_1 = \theta/2$, we obtain

$$EX_i = \left(1 - \frac{1}{2^i}\right)\theta, \quad i = 1, 2, \ldots, n. \tag{2.2}$$

Define

$$u_i = \left(1 - \frac{1}{2^i}\right), \quad i = 1, 2, \ldots, n. \tag{2.3}$$

Similarly, we obtain for $1 \leqslant i \leqslant j \leqslant n$

$$\operatorname{cov}(X_i, X_j) = \left(\frac{1}{2^{j-i}3^i} - \frac{1}{2^{i+j}}\right)\theta^2. \tag{2.4}$$

Define for $1 \leqslant i \leqslant j \leqslant n$

$$\sigma_{ij} = \left(\frac{1}{2^{j-i}3^i} - \frac{1}{2^{i+j}}\right), \tag{2.5}$$

and let Σ be the $n \times n$ matrix whose ij^{th} element is σ_{ij}.

3. ESTIMATING THE VALUE θ FOR A FIXED SAMPLE SIZE

We now treat θ as a parameter unknown to the statistician. We will compare three approaches for estimating θ based on a sample of n bids.

3.1 Functions of the Maximum Likelihood Estimate

The maximum bid X_n is the MLE for θ. Since X_n is not sufficient for θ, we would expect to do better with estimators based on more information. We will, however, consider efficiency questions at the end of this section for two estimators based on the MLE. Let $\hat{\theta}_1$ denote the unbiased function of the MLE given by

$$\hat{\theta}_1 = X_n/u_n, \tag{3.1}$$

where u_i is defined in (2.3). Recall from (2.5) that

$$\sigma_{nn} = \frac{1}{3^n} - \frac{1}{4^n}. \tag{3.2}$$

The variance (or mean-squared error) of the estimate $\hat{\theta}_1$ may be shown to be

$$\operatorname{var}(\hat{\theta}_1) = \frac{\sigma_{nn}\theta^2}{u_n^2}. \tag{3.3}$$

Let $\hat{\theta}_2$ be the multiple of the MLE with smallest mean-squared error. One may easily obtain that

$$\hat{\theta}_2 = \left(\frac{u_n}{u_n^2 + \sigma_{nn}}\right) X_n. \tag{3.4}$$

The mean-squared error of the estimate $\hat{\theta}_2$ may be shown to be

$$MSE(\hat{\theta}_2) = \left(\frac{\sigma_{nn}}{u_n^2 + \sigma_{nn}}\right)\theta^2. \tag{3.5}$$

3.2. The Best Linear Unbiased Estimate

We now derive the best linear unbiased estimate of θ, that is, the estimate of the form

(3.6) $$\hat{\theta}_a = \mathbf{a}^T\mathbf{X}$$

with minimum variance under the constraint

(3.7) $$\mathbf{a}^T E\mathbf{X} = \theta,$$

which may be rewritten as

(3.8) $$\mathbf{a}^T\mathbf{u} = 1,$$

The BLUE derived in this paragraph is an improvement over $\hat{\theta}_1$ since $\hat{\theta}_1$ is itself a linear unbiased estimate of θ. The relative efficiency of $\hat{\theta}_2$ with respect to the BLUE is discussed at the end of section 3. The variance of the estimator $\hat{\theta}_a$ is given by

(3.9) $$\operatorname{var}(\hat{\theta}_a) = \theta^2\mathbf{a}^T\Sigma\mathbf{a}$$

which is to be minimized over vectors $\mathbf{a}$ satisfying (3.8). Using straight-forward Lagrange minimization, we have by differentiation

(3.10) $$2\theta^2\Sigma\mathbf{a} = \lambda\mathbf{u}.$$

Since $\mathbf{a}$ must satisfy (3.8), we may identify λ as

(3.11) $$\lambda = \frac{2\theta^2}{\mathbf{u}^T\Sigma^{-1}\mathbf{u}}$$

so that the BLUE, which we denote by $\hat{\theta}_3$, is given by

(3.12) $$\hat{\theta}_3 = \frac{\mathbf{u}^T\Sigma^{-1}\mathbf{X}}{\mathbf{u}^T\Sigma^{-1}\mathbf{u}}.$$

The variance of $\hat{\theta}_3$ is given by

(3.13) $$\operatorname{var}(\hat{\theta}_3) = \frac{\theta^2}{u^T\Sigma^{-1}u}.$$

3.3 A Best Invariant Estimate

We note that θ is a scale parameter of the distribution of $\mathbf{X}$, that is, the density f given in (2.1) may be written in the form

$$f(x_1, \ldots, x_n|\theta) = \left(\frac{1}{\theta}\right)^n g\left(\frac{x_1}{\theta}, \ldots, \frac{x_n}{\theta}\right)$$

for $0 < \frac{x_1}{\theta} < \frac{x_2}{\theta} < \ldots < \frac{x_n}{\theta} < 1$. We develop in the paragraphs that follow a Pitman-type estimate of θ in a context to be explained shortly. We use the notation and approach of Ferguson [2, pp. 186-190] in this development.

We propose as a loss criterion the scale-invariant loss function

(3.14) $$L(\theta, a) = \left(\frac{a}{\theta} - 1\right)^2.$$

With this loss function, the decision problem of estimating θ is invariant under the group of scale transformations. Invariant decision rules are those rules d which satisfy for all $c \in \mathrm{IR}^+$

(3.15) $$d(cX_1, cX_2, \ldots, cX_n) = cd(X_1, \ldots, X_n).$$

Let

$$Y_i = \frac{X_i}{X_1} \quad \text{for } i = 2, \ldots, n. \tag{3.16}$$

The distribution of $(Y_2, \ldots, Y_n)$ does not depend on θ. We will find the best invariant estimate of θ in the conditional problem based on $f(X_1 | Y_2, \ldots, Y_n, \theta)$, where θ is still a scale parameter. We are able to establish by the usual argument that this rule is best invariant in the original problem based on $f(X_1, \ldots, X_n|\theta)$. An invariant estimate in the conditional problem has the form

$$d_{\mathbf{Y}}(X_1) = b_0 X_1,$$

where $b_0 = b_0(\mathbf{Y})$ is such that

$$E\left[\left(\frac{bX_1}{\theta} - 1\right)^2 | \mathbf{Y}\right] \tag{3.17}$$

is minimized by $b = b_0$.

The transformation T: $(X_1, X_2, \ldots, X_n) \rightarrow (X_1, Y_2, \ldots, Y_n)$ has Jacobian

$$|J| = X_1^{n-1}, \tag{3.18}$$

and thus

$$f(x_1, y_2, \ldots, y_n|\theta) = \begin{cases} x_1^{n-1} \prod_{i=0}^{n-1} \dfrac{1}{\theta - y_i x_1} & \text{for } 1 < y_2 < \ldots < y_n < \infty \\ & \text{and } 0 < x_1 < \dfrac{\theta}{y_n} \\ 0 & \text{otherwise,} \end{cases} \tag{3.19}$$

where we take $y_0 \equiv 0$ and $y_1 \equiv 1$. Since the conditional density $f(x_1|\mathbf{y}, \theta)$ does not come into play in our calculations, it will not be derived. Since the risk function (3.17) of each invariant rule is a constant independent of θ, we will take $\theta = 1$ for convenience. The conditional risk at $\theta = 1$ for the rule $d_{\mathbf{Y}}(X) = bX_1$ is

$$\frac{1}{f(y_2, \ldots, y_n)} \int_0^{1/y_n} \frac{(by - 1)^2 y^{n-1}}{\prod_{i=1}^{n-1} (1 - y_i y)}\, dy, \tag{3.20}$$

where x_1 has been replaced by the dummy variable y. Minimizing (3.20) with respect to b yields

$$b_0 = \frac{\int_0^{1/y_n} \dfrac{y^n}{\prod_{i=1}^{n-1} (1 - y_i y)}\, dy}{\int_0^{1/y_n} \dfrac{y^{n+1}}{\prod_{i=1}^{n-1} (1 - y_i y)}\, dy} \tag{3.21}$$

We now proceed with the evaluation of b_0.

Consider the numerator in (3.21). The decomposition of the integrand into partial fractions yields the following representation:

$$\int_0^{1/y_n} \frac{y^n}{\prod_{i=1}^{n-1} (1 - y_i y)}\, dy = \int_0^{1/y_n} y^n \left[\sum_{i=1}^{n-1} \frac{A_i}{1 - y_i y}\right] dy \tag{3.22}$$

where $A_1, \ldots, A_{n-1}$ are the solutions of the system of linear equations

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ a_{11} & a_{12} & \cdots & a_{1,n-1} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{n-2,1} & a_{n-2,2} & \cdots & a_{n-2,n-1} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ A_{n-1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \tag{3.23}$$

where

$$a_{ij} = \sum_{\substack{1 \leqslant k_1 < \cdots < k_i \leqslant n-1 \\ k_r \neq j \text{ for all } r}} \left(\prod_{r=1}^{i} y_{k_r} \right). \tag{3.24}$$

The elements $\{a_{ij} | i = 1, \ldots, n-2\}$ in the j^{th} column of the $(n-1) \times (n-1)$ matrix in (3.23) are simply the elementary symmetric functions in the letters $y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_{n-1}$. We evaluate the integrals in (3.22) by repeated applications of the recursive formula

$$\int \frac{x^m}{a + bx}\, dx = \frac{1}{bm}\, x^m - \frac{a}{b} \int \frac{x^{m-1}}{a + bx}\, dx. \tag{3.25}$$

We thus obtain for any positive integer k

$$\begin{aligned} &\int_0^{1/y_n} \frac{y^k}{1 - y_i y}\, dy \qquad (3.26) \\ &= \frac{1}{y_i} \frac{1}{k} \left(\frac{1}{y_n} \right)^k + \frac{1}{y_i} \int_0^{1/y_n} \frac{y^{k-1}}{1 - y_i y}\, dy \\ &= - \sum_{j=1}^{k} \left(\frac{1}{y_i} \right)^j \frac{1}{k - j + 1} \left(\frac{1}{y_n} \right)^{k-j+1} + \left(\frac{1}{y_i} \right)^k \int_0^{1/y_n} \frac{dy}{1 - y_i y} \\ &= - \left[\sum_{j=1}^{k} \left(\frac{1}{y_i} \right)^j \frac{1}{k - j + 1} \left(\frac{1}{y_n} \right)^{k-j+1} + \left(\frac{1}{y_i} \right)^{k+1} \ln \left(1 - \frac{y_i}{y_n} \right) \right]. \end{aligned}$$

The integral above can be indexed by the upper limit of integration, y_n^{-1}, the exponent of the numerator, k, and the bid ratio involved, y_i. We thus denote the expression in (3.26) by $I(y_n^{-1}, k, y_i)$. Recalling (3.21) and (3.22), we may write the best invariant estimate of θ in the conditional problem as

$$\hat{\theta}_4 = \left(\frac{\sum_{i=1}^{n-1} A_i I(y_n^{-1}, n, y_i)}{\sum_{i=1}^{n-1} A_i I(y_n^{-1}, n+1, y_i)} \right) X_1. \tag{3.27}$$

The optimality of $\hat{\theta}_4$ among invariant estimates in the unconditional problem may be established by the usual argument (see, for example, [2, p. 187]).

3.4 A Comparison of $\hat{\theta}_i$, $i = 1, 2, 3, 4$

It is easy to demonstrate the inadmissibility of $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ under any loss function of the form

$$L(\theta,a) = w(\theta)(\theta - a)^2,$$

which of course includes squared error loss as well as the loss function of (3.14). The risk function of each of these estimators is dominated (uniformly in θ) by the risk function of $\hat{\theta}_4$. One may argue as follows. The estimator $\hat{\theta}_4$ is uniquely best invariant (up to equivalence) under the loss criterion (3.14). Moreover, $\hat{\theta}_i$, for $i = 1, 2, 3$, are scale invariant estimators of θ. Thus, for all θ,

$$E\left\{\frac{\hat{\theta}_4}{\theta} - 1\right\}^2 < E\left\{\frac{\hat{\theta}_i}{\theta} - 1\right\}^2, \tag{3.28}$$

for $i = 1, 2, 3$. Multiplying both sides of (3.28) by $w(\theta)\theta^2$ establishes the claim.

It remains to examine the efficiency of the estimates $\hat{\theta}_i$, $i = 1, 2, 3$, among each other as well as relative to $\hat{\theta}_4$. In other words, we week to answer the question: "How much is lost by using an alternative to the very imposing estimator $\hat{\theta}_4$?" In Table 1, we present exact relative efficiencies for $\hat{\theta}_1$, $\hat{\theta}_2$ with respect to $\hat{\theta}_3$, that is, we list the ratios of their respective mean-squared errors. It is interesting to note that $\hat{\theta}_2$ is to be preferred to $\hat{\theta}_3$ for a sample of two bids, even through it rapidly loses ground as the sample site increases.

TABLE 1

n	$RE(\hat{\theta}_1/\hat{\theta}_3)$	$RE(\hat{\theta}_2/\hat{\theta}_3)$
2	0.964	1.048
3	0.917	0.943
4	0.868	0.876
5	0.824	0.826
6	0.787	0.788
7	0.758	0.758
8	0.735	0.735
9	0.718	0.718
10	0.705	0.705

The mean-squared error of $\hat{\theta}_4$ is difficult to obtain analytically. We have generated 500 samples of sizes 2, 3, and 4 from the uniform auction with $\theta = 10$. $\mathrm{MSE}(\hat{\theta}_3)$ and $\mathrm{MSE}(\hat{\theta}_4)$ were simultaneously estimated from these samples, and the estimated efficiency of $\hat{\theta}_3$ relative to $\hat{\theta}_4$ appears in Table 2.

TABLE 2

n	$RE(\hat{\theta}_3/\hat{\theta}_4)$
2	0.922
3	0.952
4	0.981

It seems reasonable to conjecture even from this very small simulation study that the BLUE performs adequately for samples of size $n \geq 3$ and clearly provides substantial simplification in the computations required. Combining Tables 1 and 2 for $n = 2$, one may estimate $RE(\hat{\theta}_2/\hat{\theta}_4)$ as 0.966.

4. A BAYESIAN FORMULATION OF THE ESTIMATION PROBLEM

In the last section, estimators of θ were developed based on a sample of fixed size n from the uniform auction. While there are situations in which one might reasonably ignore the stop-

ping mechanism in an auction, a general treatment of the estimation problem should allow for the influence of random stopping. The estimation problem is perhaps best formulated as a sequential decision problem, a formulation in which the stopping rule is dealt with explicitly. We examine such a formulation in this section. Sequential decision problems are discussed in detail by Ferguson [2, Chapter 7].

A stopping rule in a sequential decision problem is a sequence of functions $\Phi(\mathbf{x}) = [\phi_0, \phi_1(x_1), \phi_2(x_1, x_2), \ldots]$, where $\phi_n(x_1, \ldots, x_n)$ represents the probability that sampling is terminated after n observations, given that $X_1 = x_1, \ldots, X_n = x_n$ were observed. A terminal decision rule is a sequence of functions $d(\mathbf{x}) = [d_0, d_1(x_1), d_2(x_1, x_2), \ldots]$, where d_n represents the decision rule to be employed if sampling stops after n observations. In many sequential problems, the statistician is able to select Φ as well as d, so that the focus is on choosing the pair (Φ, d) optimally. The auction setting we study here is such that Φ is not completely controllable (for example, bidding may stop after one high bid) and is generally unknown. We therefore are primarily interested in the selection of optimal estimators for given fixed stopping rules.

It is well known that the Bayesian approach to sequential estimation results in estimators that are optimal (in the Bayes sense) regardless of what stopping rule is in force, providing the stopping rule does not depend on θ. A precise statement of this result may be found in Ferguson [2, page 314]. Thus, this approach is well suited to situations in which the stopping rule is either unknown or beyond the control of the statistician. The more general (and substantially more difficult) sequential problem with stopping rules which may depend on θ will not be treated here. We develop below a class of Bayes estimators relative to Pareto prior distributions on θ and squared error loss based on a sample of n bids from the uniform auction. The sequential decision procedure which minimizes the Bayes risk for any given stopping rule Φ is the procedure which uses the Bayes estimator for the sample that becomes available, regardless of size.

The Pareto density with parameters r and θ_0 is given by

$$g(\theta|r, \theta_0) = \begin{cases} \dfrac{r\theta_0^r}{\theta^{r+1}} & \text{if } \theta \geqslant \theta_0, \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

where $\theta_0 > 0$ and $r > 0$. If a random variable θ has a Pareto distribution with $r > 1$, then

$$E\theta = \frac{r\theta_0}{r-1},$$

and if $r > 2$,

$$\operatorname{var}(\theta) = \frac{r\theta_0^2}{(r-1)^2(r-2)}.$$

We derive in this section an expression for the Bayes estimate of θ with respect to Pareto prior distributions. The expression (4.2) involves definite integrals that are difficult to evaluate in general. We evaluate the Bayes estimate only for positive integer values of the Pareto parameter r, but indicate a simple transformation which is helpful in evaluating the estimate for positive rational r.

The density function of the vector of bids in the uniform auction is given in (2.1). Thus, taking θ as a Pareto random variable with parameters r and θ_0, we obtain a posterior density proportional to

$$\frac{1}{\theta^{r+2}} \prod_{i=1}^{n-1} \frac{1}{\theta - x_i}$$

for $0 < x_1 < \ldots < x_n < \theta$, and $\theta > \theta_0$. Letting $w_n = \max(x_n, \theta_0)$, we may write the Bayes estimate of θ for squared error loss as

$$\hat{\theta}_5 = \frac{\int_{w_n}^{\infty} \frac{1}{\theta^{r+1}} \prod_{i=1}^{n-1} \frac{1}{\theta - x_i} \, d\theta}{\int_{w_n}^{\infty} \frac{1}{\theta^{r+2}} \prod_{i=1}^{n-1} \frac{1}{\theta - x_i} \, d\theta}. \tag{4.2}$$

The evaluation of $\hat{\theta}_5$ involves steps quite similar to those followed in evaluating $\hat{\theta}_4$. We thus include only those details needed in clarifying the notation used in our final expression for $\hat{\theta}_5$. We first obtain a partial fraction decomposition,

$$\prod_{i=1}^{n-1} \frac{1}{\theta - x_i} = \sum_{i=1}^{n-1} \frac{B_i}{\theta - x_i}, \tag{4.3}$$

where $B_1, \ldots, B_{n-1}$ are the solutions of the system of linear equations

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ b_{11} & b_{12} & \cdots & b_{1,n-1} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ b_{n-2,1} & b_{n-2,2} & \cdots & b_{n-2,n-1} \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ B_{n-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}, \tag{4.4}$$

with

$$b_{ij} = \sum_{\substack{1 \leqslant k_1 < \cdots < k_t \leqslant n-1 \\ k_r \neq j \text{ for all } r}} \left(\prod_{r=1}^{t} x_{k_r} \right). \tag{4.5}$$

We then have the representation for the numerator of (4.2):

$$\int_{w_n}^{\infty} \frac{1}{\theta^{r+1}} \prod_{i=1}^{n-1} \frac{1}{\theta - x_i} \, d\theta = \sum_{i=1}^{n-1} B_i \int_{w_n}^{\infty} \frac{d\theta}{\theta^{r+1}(\theta - x_i)}. \tag{4.6}$$

Integrals of the type in (4.6) were studied by Chebyshev, whose famous result implies that such integrals are elementary functions whenever r is a rational number (see Ritt [10]). We develop here the exact expression for $\hat{\theta}_5$ only for positive integers r; if r is the ratio of integers s/t, however, the transformation $\theta = \lambda^t$ reduces integrals in (4.6) to integrals of rational functions of λ. The integral of a rational function can always be obtained in finite terms by the method of partial functions.

Using the recursive formula

$$\int \frac{d\theta}{\theta^m(\theta - x_i)} = \frac{1}{(m-1)x_i\theta^{m-1}} + \frac{1}{x_i} \int \frac{d\theta}{\theta^{m-1}(\theta - x_i)} + C \tag{4.7}$$

valid for $m > 1$, we obtain for any positive integer m

$$C(x_i, m) \equiv \int_{w_n}^{\infty} \frac{d\theta}{\theta^m(\theta - x_i)} \tag{4.8}$$

$$= \sum_{j=1}^{m-1} \frac{1}{(m-j)x_i^j w_n^{m-j}} + \frac{1}{x_i^m} \ln \left[\frac{w_n - x_i}{w_n} \right]$$

We may thus express the Bayes rule relative to the Pareto prior with parameters $r \in Z^+$ and θ_0 as

$$\hat{\theta}_5 = \frac{\sum_{i=1}^{n-1} B_i C(x_i, r+1)}{\sum_{i=1}^{n-1} B_i C(x_i, r+2)} \tag{4.9}$$

and a similar expression may be developed when r is a positive rational.

We close this section with two remarks. First, we note that Bayes estimates for the Pareto prior and certain weighted, squared-error loss functions can be developed in a similar fashion. In particular, the Bayes estimate relative to the Pareto prior, with parameters $r \in Z^+$ and θ_0 and the loss function of (3.14), is given by

$$\hat{\theta}_6 = \frac{\sum_{i=1}^{n-1} B_i C(x_i, r+3)}{\sum_{i=1}^{n-1} B_i C(x_i, r+4)}.$$

Secondly, for positive integers r, one may obtain an interesting class of limits of Bayes rules as $\theta_0 \to 0$. The form of these rules is easily obtained — one simply replaces w_n in the expression for $\hat{\theta}_5$ or $\hat{\theta}_6$ by x_n. One may show that these limits of Bayes rules are invariant under scale transformations, indicating that the invariant procedures studied in section 3 are similar to Bayes procedures for suitable choices of Pareto parameters.

ACKNOWLEDGMENTS

The authors wish to thank George T. Duncan for several ideas he graciously contributed to the study of the estimation problem we have described. A helpful conversation with S. Zacks is also acknowledged.

REFERENCES

[1] Christenson, Charles, "Strategic Aspects of Competitive Bidding for Corporate Securities," Ph.D. Dissertation, Graduate School of Business Administration, Harvard University, Boston, Massachusetts (1965).

[2] Ferguson, Thomas S., *Mathematical Statistics: A Decision Theoretic Approach*, (Academic Press, New York, 1967).

[3] Gaver, Donald P., "Random Record Models" (abstract), Bulletin of the Institute of Mathematical Statistics *5*, 73 (1976).

[4] Griesmer, J. H., and M. Shubik, "Toward a Study of Bidding Processes: Some Constant-Sum Games," Naval Research Logistics Quarterly *10*, 11-21 (1963).

[5] Griesmer, J. H., and M. Shubik, "Toward a Study of Bidding Processes, Part II: Games with Capacity Limitations," Naval Research Logistics Quarterly *10*, 151-173 (1963).

[6] Griesmer, J. H., and M. Shubik, "Toward a Study of Bidding Processes, Part III: Some Special Models," Naval Research Logistics Quarterly *10*, 199-217 (1963).

[7] Griesmer, J. H., M. Shubik, and R. E. Levitan, "Toward a Study of Bidding Processes, Part IV: Games with Unknown Costs," Naval Research Logistics Quarterly *14*, 415-433 (1967).

[8] Lehmann, E. L., and H. Scheffe, "Completeness, Similar Regions and Unbiased Estimation, Part I," Sankhya *10*, 305-340 (1950).

[9] Pitman, E. J. G., "The Estimation of Location and Scale Parameters of a Continuous Population of Any Given Form," Biometrika *30*, 391-421 (1939).
[10] Ritt, J. F., *Integration in Finite Terms* (Columbia University Press, New York, 1948).
[11] Sosnick, S. H., "On the Distribution of Price Changes at a Bull Auction," Journal of Farm Economics *47*, 1306-1320 (1965).
[12] Stark, Robert M., "Competitive Bidding: A Comprehensive Bibliography," Operations Research *19*, 484-490 (1971).
[13] Stein, Sherman K., *Calculus and Analytic Geometry* (McGraw-Hill, New York, 1973).
[14] Vickrey, William, "Counterspeculation, Auctions and Competitive Sealed Tenders," Journal of Finance *16*, 8-37 (1961).
[15] Yang, M., "On the Distribution of the Inter-Record Times in an Increasing Population," Journal of Applied Probability *12*, 148-154 (1975).

THE OPTIMUM DESIGN OF MULTIVARIATE ACCEPTANCE SAMPLING PLANS

S. C. Chapman and J. W. Schmidt

Department of Industrial Engineering and Operations Research
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

G. K. Bennett

Department of Industrial Systems
University of South Florida
Tampa, Florida

ABSTRACT

A model is developed which may be used to determine the expected total cost of quality control per inspection lot under acceptance sampling by variables where several characteristics are to be simultaneously controlled. Optimization of the model is accomplished through the application of a conventional search procedure. The sensitivity of the model and the optimum solution to the shape of the underlying probability distributions is discussed and associated analyses are presented through an example.

INTRODUCTION

Over the past decade considerable attention has been devoted to the subject of cost-based quality-control systems. Hald [12] and Horsnell [15] have developed and analyzed mathematical cost models for acceptance sampling for a single attribute while Duncan [5, 7], Gibra [10], and Goel, et al., [11] have carried out similar research with respect to process control of a single variable. Relatively little attention has been given to the problem of simultaneous control of several attributes or variables. Schmidt and Bennett [29] have developed a cost-based model for multi-attribute acceptance sampling. Ghare and Torgersen [9] have examined the simultaneous control of several variables where the basis for control is purely statistical. Montgomery and Klatt [20] have extended the Hotelling T^2 Model to include economic considerations. However, the quality-control system resulting from the model by Montgomery and Klatt identifies an out-of-control situation but does not specify the variable or variables causing this condition. Latimer, Bennett, and Schmidt [18] have presented a model for multivariate process control. However, an analogous model for cost-based acceptance sampling is not available. This is the topic of this paper.

THE SYSTEM MODELED

The problem addressed in this paper is that of multivariate acceptance sampling where the quality characteristics of concern are independently distributed random variables. The mathematical model developed measures the expected total cost of quality control per

inspection lot where the sample mean for each characteristic is the statistic used to assess the acceptability of the lot with respect to that characteristic. The decision variables included in the model are the sample size and lower and upper acceptance limits for the sample mean for each characteristic. The model applies to the case where inspection may be destructive or nondestructive; rejection of the inspection lot on some characteristics results in scrapping of the lot, while screening results when rejection occurs on others. The model developed takes account of the costs associated with inspection, with screening and repair or replacement of defective items when the lot is rejected and screened, with scrapping the lot when the lot is rejected and scrapped, and with defects present in accepted and screened lots.

The model presented here is appropriate in cases where the acceptability of manufactured-product lots is dependent upon the lot mean values of several measurable characteristics. The lot means for the several characteristics are assumed to be independently distributed random variables, and the mechanism used to assess lot acceptability is a variables quality-control system for the lot mean of each characteristic. The problem is a familiar one in the manufacturing industry, although the characteristic lot means are frequently correlated. However, cases where these characteristics are independently distributed are not uncommon, as illustrated by the following example.

Consider a process for the production of collapsible metal tubes, such as tooth-paste tubes. These tubes are manufactured on an automated line in lots of approximately 14,000. Four characteristics are of particular concern: tube diameter, tube length, neck length, and cap torque. Tube diameter is controlled by an extrusion process which results in a formed tube about 1 in. longer than required. The end of the tube is then cut off, the final length being controlled by measurement from the neck of the tube. Plastic necks are then automatically inserted in the top of the tube; and the projection of the neck is controlled by the insertion process. The final operation is capping the tube, and the torque required to remove the cap is dependent upon proper adjustment of the capping operation. Analyses to date have indicated that there is no significant correlation among these variables for the present process, although the process used previously resulted in slight correlation between tube diameter and length. The latter condition resulted when the extrusion process produced tubes which were shorter than the desired final length. However, this failure has not occured on the present process.

The model may appproximate reality even when mild correlation exists among the characteristics of concern. However, this conjecture cannot be verified at this time, since cost models for the case of correlated characteristics have not been addressed in the literature. Hence, the model developed here represents only a first step in the analysis of multivariate acceptance sampling for quality control.

The model assumes that the characteristics of concern may be classified into one of three mutually exclusive categories. The first category, characteristic class I, includes those variables for which inspection is destructive to the extent that the product is rendered unfit for its intended use and for further inspection on other characteristics. The second category, characteristic class II, includes those variables for which inspection is destructive in the sense that the product cannot be used for its intended purpose once inspection on these characteristics takes place. However, inspection on characteristics in class II does not render the product unfit for further inspection. The third category, characteristic class III, includes those variables for which inspection is nondestructive. As the above discussion implies, rejection of a lot on variables falling into classes I or II results in scrapping of the inspection lot, while rejection on variables in class III only leads to screening of the inspection lot.

Let $n_1 \leqslant n_2 \leqslant \ldots \leqslant n_{k_1}$ be the sample sizes for variables 1 through k_1 in class I, $n_{k_1+1} \leqslant n_{k_1+2} \leqslant \ldots \leqslant n_{k_2}$ the sample sizes for variables $k_1 + 1$ through k_2 in class II, and $n_{k_2+1} \leqslant \ldots \leqslant n_{k_3}$ the sample sizes for variables $k_2 + 1$ through k_3 in class III. The inspection procedure is sequential and can be summarized by the decision tree in Figure 1, where the notation $I[k, l, S_l(N)]$ indicates inspection of k units on characteristic l and the k units inspected are drawn from set S_l, which includes N units.

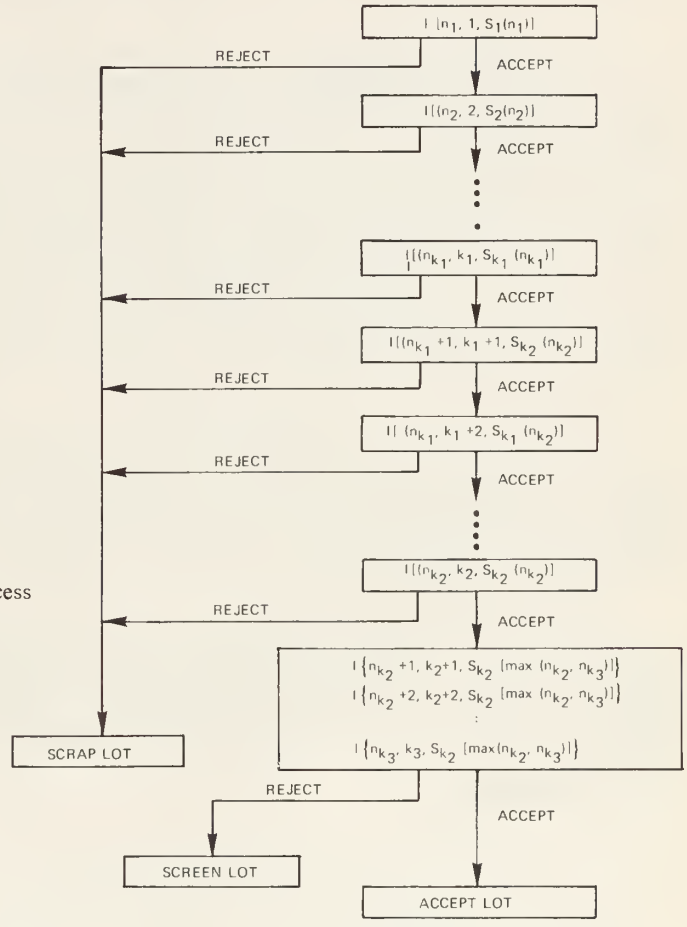

FIGURE 1. Decision tree for the inspection process

The costs arising from the implementation of a multivariate acceptance sampling plan may be considered to fall into two categories: the costs incurred in reaching a decision to accept or reject the inspection lot, and those costs arising from implementation of the decision to accept or reject the lot. The cost of reaching the decision to accept or reject a lot is the cost of inspecting the units included in the random sample drawn from the inspection lot. Specifically, if C_{I_i} is the cost of inspecting one unit for the i th characteristic, and the number of units inspected is n_i, then the cost of inspection for the i th variable is $C_{I_i}\, n_i$.

When an inspection lot is accepted, several cost components must be considered. First, those units destroyed in the inspection process result in a loss of product value and a materials-handling cost, and they may be replaced, resulting in an additional cost of replacing the destroyed units. Secondly, those units in the sample which are not destroyed during inspection but are found to be defective as the result of nondestructive inspection may be repaired. Finally, one cannot assume that an accepted inspection lot is defect free. Thus, it is assumed that a cost is associated with each characteristic defect present in an accepted inspection lot.

If the inspection lot is rejected and scrapped, the cost incurred is that resulting from loss the investment in the lot as well as the cost of materials handling and carrying out the scrapping operation. When the inspection lot is rejected and screened, the resulting costs are those of screening the units included in the uninspected portion of the lot and in the nondestroyed portion of the sample on the variable or variables for which screening is called, the cost of repairing defective units found during screening, the cost of discarding and possibly replacing destroyed units, and the cost of units which are defective with respect to those variables for which lot acceptance occurred.

The random variables considered in the system are the lot mean, the sample mean, and the individual unit dimension for each variable of concern. If S_{Li} and S_{Ui} are the lower and upper specification limits for the ith variable, and if X_i is the measured value of that variable for a given unit of product, then that unit of product is considered defective with respect to the ith variable if $X_i < S_{Li}$ or $X_i > S_{Ui}$. Otherwise the product is acceptable with respect to the ith variable. Acceptance or rejection of an inspection lot with respect to a given variable is determined by the value of sample mean for that variable and the associated acceptance limits for the sample mean. Specifically, if LCL_i and UCL_i are the lower and upper acceptance limits for the sample mean for the i th variable, and if $\bar{X}_i$ is the value of the sample mean for that variable, then the inspection lot is accepted with respect to the ith variable if $LCL_i \leqslant \bar{X}_i \leqslant UCL_i$ and is rejected otherwise.

The purpose of the multivariate quality-control system modeled here is to detect, isolate, and, in the case of screening, rectify those inspection lots where the lot mean for one or more variables significantly deviates from its desired value. Since one cannot normally predict the value of the lot mean for a given variable with certainty, the lot mean is considered to be a random variable, as indicated above.

The model developed in this paper represents the expected total cost of quality control per inspection lot and is a function of the decision variables mentioned above. The values of the decision variables which minimize the expected total cost of quality control per lot are determined through the use of a conventional search technique, the pattern search [42].

Additional Notation

L = the number of items in an inspection lot.
(μ_i, σ_i^2) = mean and variance of product dimension for the i th variable in an inspection lot.
$\vec{\mu}$ = the vector $(\mu_1, \mu_2, \ldots, \mu_{k_3})$.
$f(\mu_i)$ = probability density function of μ_i.
$g(\bar{x}_i|\mu_i)$ = conditional probability density function of $\bar{X}_i$ for samples drawn from an inspection lot with mean μ_i.
$h(x_i|\mu_i)$ = conditional probability density function of X_i for units drawn from an inspection lot with mean μ_i.

$$P(D_{Li}|\mu_i) = \int_{-\infty}^{S_{Li}} h(x_i|\mu_i)\, dx_i.$$

$$P(D_{Ui}|\mu_i) = \int_{S_{Ui}}^{\infty} h(x_i|\mu_i)\, dx_i.$$

$$P(G_i|\mu_i) = 1 - P(D_{Li}|\mu_i) - P(D_{Ui}|\mu_i).$$

$$P(A_i|\mu_i) = \int_{LCL_i}^{UCL_i} g(\bar{x}_i|\mu_i)\, d\bar{x}_i.$$

$C_I(\vec{\mu})$ = expected cost of sampling inspection given the vector $\vec{\mu}$.

C_{aLi}, C_{aUi} = cost which results from the presence of the i th variable below the lower and above the upper specification limit in an accepted lot.

C_A (accept, $\vec{\mu}$) = expected cost of accepting the inspection lot given acceptance and the vector $\vec{\mu}$.

$C_A(\vec{\mu})$ = expected cost of accepting the inspection lot given the vector $\vec{\mu}$.

C_p, C_r = unit cost of production and materials handling respectively.

C_R (scrap, $\vec{\mu}$) = expected cost of scrapping the inspection lot given rejection and the vector $\vec{\mu}$.

$C_R(\vec{\mu})$ = expected cost of scrapping the inspection lot given the vector $\vec{\mu}$.

C_{si} = unit cost of performing the screening inspection for the i th variable.

$C_{\sigma Li}$, $C_{\sigma Ui}$ = cost of repairing one unit of product found to have the i th variable below the lower and above the upper specification limits respectively.

$C_S(\vec{\mu})$ = expected cost of screening the inspection lot given the vector $\vec{\mu}$.

C_T = total expected cost of quality control per inspection lot submitted for control.

N^* = total number of items removed from the inspection lot for the purpose of sampling inspection.

$$= \max\left\{\sum_{i=1}^{k_1} n_i + n_{k_2},\ \sum_{i=1}^{k_1} n_i + n_{k_3}\right\}$$

$$N_{1i} = \min\{n_i - n_{i-1},\ n_i - n_{k_2}\},\ i = k_2 + 1,\ \ldots,\ k_3$$

$$N_{2i} = \max\{n_i - n_{k_2},\ 0\},\ i = k_2 + 1,\ \ldots\ k_3$$

$$\delta_0 = \begin{cases} 1, & n_{k_3} > n_{k_2} \\ 0, & n_{k_3} \leqslant n_{k_2} \end{cases}$$

$$\delta_{1i} = \begin{cases} 1, & n_i > n_{k_2} \\ 0, & n_i \leqslant n_{k_2} \end{cases} \qquad i = k_2 + ,\ \ldots,\ k_3$$

$$\delta_2 = \begin{cases} 1, \text{ units destroyed during sampling inspection are replaced in accepted or screened lots} \\ 0, \text{ units destroyed during sampling inspection are not replaced in accepted or screened lots} \end{cases}$$

Additional Assumptions

1. The random variables $\mu_1, \ldots, \mu_{k_3}$ are independently distributed, as are $\bar{X}_1, \ldots, \bar{X}_{k_3}$ and $X_1, \ldots, X_{k_3}$.

2. Errors, if any, which occur in sampling or screening inspection are assumed to be negligible.

3. A single defect type can occur only once in a unit of product.

4. A unit that has been rendered unfit for its intended use during sampling inspection is scrapped, but it may or may not be replaced if the inspection lot is accepted or screened.

5. In an inspeciton lot that has not been scrapped, a unit that has not been destroyed during sampling inspection but has been found defective during subsequent sampling or screening inspection on a nondestructible variable type is assumed to be repaired.

6. A unit that has been accepted, or repaired, but has multiple defects will incur the cumulative cost of accepting, or repairing, each characteristic type.

7. If the inspection lot is accepted on all destructible characteristic types, and the policy is to replace all units destroyed in sampling inspection, replacement units are screened on all nondestructible characteristic types and defectives are repaired accordingly.

8. When screening for multiple characteristic types, the inspection of any one unit is carried out to completion even though a defect in the unit may be detected prior to the inspection of all characteristic types on that unit.

9. An inspection lot which is accepted on all destructible characteristic types cannot be scrapped. If the inspection lot is subsequently rejected on one or more screenable characteristic types, the remainder of the inspection lot is screened for those variable types only.

MODEL DEVELOPMENT

Inspection Cost

If sampling is conducted on the i th variable, the cost of inspection is $C_{Ii}n_i$. For variables in classes I and II, $i = 1, 2, \ldots, k_2$, the conditional probability that inspection is carried out, given $\vec{\mu}$, is unity for $i = 1$ and $\prod_{j=1}^{i-1} P(A_j|\mu_j)$ for $i = 2, 3, \ldots, k_2$. Inspection is carried out on variables in class III only if the inspection lot is accepted on all variables in classes I and II. Hence, the conditional probability that inspection is carried out on variables in class III, given $\vec{\mu}$, is $\prod_{j=1}^{k_2} P(A_j|\mu_j)$.

Thus

$$C_I(\vec{\mu}) = C_{I1}n_1 + \sum_{i=2}^{k_2}\left[C_{Ii}n_i \prod_{j=1}^{i-1} P(A_j|\mu_j)\right] + \left[\sum_{i=k_2+1}^{k_3} C_{Ii}n_i\right]\prod_{j=1}^{k_2} P(A_j|\mu_j). \tag{1}$$

Acceptance Cost

The first cost considered when an inspection lot is accepted is that resulting from units destroyed (classes I and II) and units repaired (class III) in the sample. Thus, the conditional expected cost of units destroyed or repaired given lot acceptance and $\vec{\mu}$ is given by

(2) E(cost of destroyed and repaired units | accept, $\vec{\mu}$),

$$= (C_\rho + C_r)\left(\sum_{i=1}^{k_1} n_i + n_{k_2}\right) + C_r\left\{\sum_{i=k_2+1}^{k_3} \delta_{1i} N_{1i}\left[1 - \prod_{j=i}^{k_3} P(G_j|\mu_j)\right]\right\}$$

$$+ \sum_{i=k_2+1}^{k_3} N_{2i}\,[C_{\sigma Ui}\,P(D_{Ui}|\mu_i) + C_{\sigma Li}P(D_{Li}|\mu_i)].$$

It is assumed that units destroyed during sampling inspection may or may not be replaced if the inspection lot is accepted. The policy to replace or not to replace is defined by δ_2. All replacement units are screened, and repaired where necessary, on class III variables. However, replacement units which are defective with respect to class I or II variables pass undetected. Hence,

(3) E(replacement cost | accept, $\vec{\mu}$) $= \delta_2\left(\sum_{j=1}^{k_1} n_j + n_{k_2}\right)\left\{C_\rho\right.$

$$+ \sum_{i=k_2+1}^{k_3} [C_{si} + C_{\sigma Ui}P(D_{Ui}|\mu_i) + C_{\sigma Li}P(D_{Li}|\mu_i)]$$

$$\left. + \sum_{i=1}^{k_2} [C_{aUi}P(D_{Ui}|\mu_i) + C_{aLi}P(D_{Li}|\mu_i)]\right\}.$$

When an inspection lot is accepted, there may be some units in the sample which have not been destroyed and which have not been inspected on some of the variables of concern. Such units may contain defects with respect to those variables for which inspection was not conducted and the associated expected cost, given acceptance and $\vec{\mu}$, is given by

(4) E(cost of defects in uninspected and nondestroyed portion of the sample | accept, $\vec{\mu}$)

$$= \sum_{i=k_2+1}^{k_3}\left\{\delta_{1i}\,N_{1i}\sum_{j=1}^{i-1}\left[C_{aUj}P(D_{Uj}|\mu_j) + C_{aLj}P(D_{Lj}|\mu_j)\right]\right\}.$$

The final cost component arising when an inspection lot is accepted is that resulting from the presence of defective units in the uninspected portion of the lot, which is comprised of L-N^* units. N^* is the total number of items removed from the lot for the purpose of sampling inspection. Then

(5) E(cost of defects in the uninspected portion of the lot | accept, $\vec{\mu}$)

$$= (L - N^*)\sum_{i=1}^{k_3}\left[C_{aUi}P(D_{Ui}|\mu_i) + C_{aLi}P(D_{Li}|\mu_i)\right],$$

The conditional expected total cost of acceptance, given acceptance and $\vec{\mu}$, is then given by

(6) $C_A(\text{accept}, \vec{\mu}) = \text{E(cost of destroyed and repaired units} \mid \text{accept}, \vec{\mu})$

$+ \text{E(replacement cost} \mid \text{accept}, \vec{\mu})$
$+ \text{E(cost of defects in uninspected and nondestroyed portion of the sample} \mid \text{accept}, \vec{\mu})$
$+ \text{E(cost of defects in the uninspected portion of the lot} \mid \text{accept}, \vec{\mu})$

and the conditional expected cost of acceptance, given $\vec{\mu}$, is

$$C_A(\vec{\mu}) = C_A(\text{accept}, \vec{\mu}) \prod_{j=1}^{k_3} P(A_j|\vec{\mu}_j). \tag{7}$$

Scrapping Cost

An inspection lot is rejected and scrapped only if the lot is rejected on one of the variables in classes I or II. The conditional expected cost of scrapping, given $\vec{\mu}$, is

$$C_R(\vec{\mu}) = L(C_r + C_\rho)\left[1 - \prod_{i=1}^{k_2} P(A_i|\mu_i)\right]. \tag{8}$$

Screening Cost

An inspection lot is screened whenever it is accepted on all variables in classes I and II but is rejected on one or more variables in class III. Screening inspection then occurs only for those variables in class III for which rejection occurs. The expected cost of carrying out the screening operation, given acceptance on class I and II variables and $\vec{\mu}$, is

(9) $\text{E(screening cost} \mid \text{acceptance on I and II variables}, \vec{\mu})$

$$= (L - N^*) \sum_{i=k_2+1}^{k_3} C_{si}[1 - P(A_i|\mu_i)].$$

In addition, the units in the sample which were not destroyed during sampling inspection and which were not inspected on the rejected variable must also be screened, if any such units exist. Hence, the total conditional cost of the screening operation, given acceptance on class I and II variables and $\vec{\mu}$, including the quantity given in (9), is

(10) $\text{E(total screening cost} \mid \text{acceptance on I and II variables}, \vec{\mu})$

$$= \sum_{i=k_2+1}^{k_3} C_{si}\left\{L - N^* + \delta_o[\delta_{1I}(n_{k_2} - n_i) + n_{k_3} - n_{k_2}]\right\}[1 - P(A_i|\mu_i)].$$

As in the case of lot acceptance, the cost of units destroyed (classes I and II) and rejected and repaired (class III) during inspection must be accounted for. The cost of units destroyed (classes I and II) or rejected and repaired (class III), given acceptance on variables in classes I and II and $\vec{\mu}$, is

(11) E(cost of destroyed and repaired units | acceptance on I and II variables, $\vec{\mu}$)

$$= \left[(C_\rho + C_r)\left(\sum_{i=1}^{k_1} n_i + n_{k_2}\right) + C_r\left\{\sum_{i=k_2+1}^{k_3} \delta_{1i}\, N_{1i}\left[1 - \prod_{j=i}^{k_3} P(G_j|\mu_j)\right]\right\} + \sum_{i=k_2+1}^{k_3} N_{2i}\,[C_{\sigma Ui}P(D_{Ui}|\mu_i) + C_{\sigma Li}P(D_{Li}|\mu_i)]\right]\left[1 - \prod_{i=k_2+1}^{k_3} P(A_i|\mu_i)\right].$$

Those units not destroyed in the sample but which have been screened and rejected on a given variable must be repaired. This cost is given by

(12) E(screening and repair cost for the sample | acceptance on class I and II variables, $\vec{\mu}$)

$$= \sum_{i=k_2+1}^{k_3}\left\{C_r\delta_{1i}(n_{k_3} - n_i)\,[1 - P(G_i|\mu_i)] + [\delta_{1i}(n_{k_3} - n_i) + (1 - \delta_{1i})\,(n_{k_3} - n_{k_2})]\right.$$

$$\left.\cdot\,[C_{\sigma Ui}P(D_{Ui}|\mu_i) + C_{\sigma Li}P(D_{Li}|\mu_i)]\,(1 - P(A_i|\mu_i))\right\}.$$

Since the screened lot is eventually treated in a fashion similar to that in the case of lot acceptance, units destroyed in the sampling inspection may or may not be replaced and, as in the case of lot acceptance, are screened on class III variables but may contain undetected defects with respect to variables in classes I and II. Hence

(13) E(replacement cost | acceptance on class I and II variables, $\vec{\mu}$)

$$= \delta_2\left(\sum_{j=1}^{k_1} n_j + n_{k_2}\right)\left\{C_\rho + \sum_{i=k_2+1}^{k_3}\left[C_{si} + C_{\sigma Ui}P(D_{Ui}|\mu_i) + C_{\sigma Li}P(D_{Li}|\mu_i)\right]\right.$$

$$\left. + \sum_{i=1}^{k_2}[C_{aUi}P(D_{Ui}|\mu_i) + C_{aLi}P(D_{Li}|\mu_i)]\right\}\left[1 - \prod_{j=k_2+1}^{k_3} P(A_j|\mu_j)\right].$$

In addition to undetected defects contained in replacement items, undetected defects with respect to a given variable may occur in units which have neither been destroyed nor inspected on that variable. This component of cost arising from screening the inspection lot can be expressed as

(14) E(undetected defects in the sample | acceptance on class I and II variables, $\vec{\mu}$)

$$= \delta_0 \sum_{i=k_2+1}^{k_3}\left\{\delta_{1i}\, N_{1i}\left[\sum_{j=1}^{k_2} C_{aUj}P(D_{Uj}|\mu_j) + C_{aLj}P(D_{Lj}|\mu_j)\right]\left[1 - \prod_{i=k_2+1}^{k_3} P(A_j|\mu_j)\right]\right.$$

$$\left. + \left[\delta_{1i}(n_{k_3} - n_i) + (1 - \delta_{1i})\,(n_{k_3} - n_{k_2})\right]\left[C_{aUi}P(D_{Ui}|\mu_i)\right.\right.$$

$$+ C_{aLi}P(D_{Li}|\mu_i)\Big] P(A_i|\mu_i) \left[1 - \prod_{\substack{j=k_2+1 \\ j\neq i}}^{k_3} P(A_j|\mu_j)\right]\Bigg\}$$

The final two components of cost associated with screening an inspection lot are those relating to the repair or replacement of defects found during screening of the uninspected portion of the lot and the occurence of undetected defects in the uninspected portion of the lot. The conditional cost of handling and repairing defects found during screening inspection of the lot, given acceptance on class I and II variables and $\vec{\mu}$, is

(15) E(retification cost for the screened lot | acceptance on class I and II variables, $\vec{\mu}$)

$$= (L - N^*) \sum_{i=k_2+1}^{k_3} \Big\{C_r[1 - P(G_i|\mu_i)] + C_{\sigma Ui}P(D_{Ui}|\mu_i) + C_{\sigma Li}P(D_{Li}|\mu_i)\Big\}$$

$$\cdot [1 - P(A_i|\mu_i)].$$

The cost of accepted defects in the uninspected portion of the lot, given acceptnce on all variables in classes I and II and $\vec{\mu}$, is

(16) E(cost of undetected defects in the uninspected portion of the lot | acceptance on class I and II variables, $\vec{\mu}$) $= (L - N^*)$

$$\Bigg\{\sum_{i=1}^{k_2} [C_{aUi}P(D_{Ui}|\mu_i) + C_{aLi}P(D_{Li}|\mu_i)] \left[1 - \prod_{i=k_2+1}^{k_3} P(A_j|\mu_j)\right]\Bigg\}$$

$$+ (L - N^*) \Bigg\{\sum_{i=k_2+1}^{k_3} [C_{aUi}P(D_{Ui}|\mu_i) + C_{aLi}P(D_{Li}|\mu_i)] P(A_i|\mu_i) \cdot \left[1 - \prod_{\substack{j=k_2+1 \\ j\neq i}}^{k_3} P(A_j|\mu_j)\right]\Bigg\}.$$

Summing (8) through (16) yields the conditional expected total cost resulting from screening, given acceptance on class I and II variables and $\vec{\mu}$, C_S (acceptance on class I and II variables, $\vec{\mu}$) and

$$(17) \qquad C_S(\vec{\mu}) = C_S(\text{acceptance on class I and II variables, } \vec{\mu}) \prod_{i=1}^{k_2} P(A_i|\mu_i).$$

The total expected cost of quality control per inspection lot, C_T, is then given by

$$(18) \qquad C_T = \prod_{i=1}^{k_3} \int_{-\infty}^{\infty} [C_I(\vec{\mu}) + C_A(\vec{\mu}) + C_R(\vec{\mu}) + C_S(\vec{\mu})] f(\mu_i)\, d\mu_i$$

RESULTS

Application of the model and its optimization are illustrated in the following example.

EXAMPLE 1: A control device is composed of three fuses, each of which controls the flow of current in a different circuit. The device is a component of a piece of electrical equipment which will fail if either too much or too little current flows in any circuit, resulting in

repair of the equipment at final testing. Inspection of the control device consists of measuring the current at which each fuse burns out. Burn out of fuses 1 or 2 normally results in damage to the device sufficient to render inspection on any other variable impractical. However, testing the third fuse does not damage the device sufficiently to eliminate testing on the other variables. A fourth variable inspected is the diameter of the device, to make sure it will fit into the final assembly. The data for this example are given in Table 1; current is measured in milliamperes and diameter is measured in 10^{-4} meters. The following approximate relations hold:

$$x_i \sim N(\mu_i, \sigma_i^2),\ \bar{x}_i \sim N\left(\mu_i, \frac{\sigma_i^2}{n_i}\right),\ \mu_i \sim N\left(\mu_{0i}, \frac{\sigma_i^2}{\gamma_i}\right),\ i = 1, 2, 3, 4.$$

TABLE 1. *Parameters for Example 1*

Variable i	Class	S_{Li}	S_{Ui}	γ_i	μ_{0i}	σ_i	C_{aUi}	C_{aLi}	$C_{\sigma Ui}$	$C_{\sigma Li}$	C_{ai}	C_{fi}
1 (fuse)	II	78.40	81.60	1.67	80.00	0.50	$55.85	$55.85	—	—	—	$0.005
2 (fuse)	II	45.80	54.20	1.67	50.00	1.40	34.75	34.75	—	—	—	0.005
3 (fuse)	II	23.00	27.00	1.67	25.00	0.80	12.00	12.00	—	—	—	0.005
4 (diameter)	III	212.30	217.70	1.67	215.00	1.00	5.00	5.00	$4.75	$4.75	$0.005	0.005
$k_1 = 2$		$k_2 = 3$		$k_3 = 4$		$L = 10{,}000$		$C_\mu = \$2.00$		$C_t = \$0.00$		$\delta_2 = 1$

Optimization of the model can be achieved through application of an iterative search procedure, in this case the pattern search. The purpose of the search is to define values of the decision variables which minimize the total cost of quality control. Because of the exploratory nature of search procedures, the "indicated" optimum may not be the true optimum. However, for those cases investigated in this research, the cost at the indicated optimum was always within 0.10% of the least-cost solution optained by subsequent enumeration about the "indicated" optimum identified by the search procedure.

Two approaches have been explored in applying an iterative search technique to optimization of the model. In both approaches a starting point representing no sampling and implying lot acceptance on all variables ($n_i = 0$, $i = 1, 2, 3, 4$) was chosen. In the first approach the pattern search is applied directly to the multivariate model with the starting point represented by lot acceptance on all variables. The second approach consists of applying the model to each variable separately, as though the remaining variables were not to be considered. The pattern search is then applied to each single-variable model, using lot acceptance as the starting point in each case. The least-cost solution for each variable treated independently is then used as the starting point for optimization of the model when all variables are considered. The latter approach has been found to be more efficient with respect to execution time on a digital computer, and it is the approach presented here.

The results of the application of the pattern search to the model when each variable is treated separately is summarized in Table 2. The components of the expected cost of quality control C_I, C_A, C_R, C_S, and C_T, where $C_I = E[C_I(\vec{\mu})]$, $C_A = E[C_A(\vec{\mu})]$, $C_R = E[C_R(\vec{\mu})]$, and $C_S = E[C_S(\vec{\mu})]$, given for each variable are those costs which arise when the variables are treated independently and are therefore misleading. As already indicated, in the second stage of the search, the least-cost solution for sample size and lower and upper control limits given in Table 2 for each variable are used as the starting point for the search for the optimum when all variables are represented in the model. The second stage of the search is summarized in Table 3. The total expected cost of quality control per inspection, C_T, and the components thereof, C_I, C_A, C_R, and C_S, for the starting point for the second stage of the search (optimum points in Table 2) and for the final optimum indicated by the search are given in Table 3.

TABLE 2. *Optimization of each variable treated independently*

Variable i	Point	n_i	LCL_i	UCL_i	C_I	C_A	C_R	C_S	C_T
1	Start	0	—	—	$0.00	$6367.62	$0.00	0.00	6367.62
	Opt.	30	79.25	80.75	0.15	3879.90	1189.01	0.00	5069.06
2	Start	0	—	—	0.00	6147.27	0.00	0.00	6147.27
	Opt.	32	47.94	52.06	0.16	3905.91	1274.14	0.00	5180.21
3	Start	0	—	—	0.00	5772.03	0.00	0.00	5772.03
	Opt.	22	23.70	26.30	0.11	4573.38	861.95	0.00	5435.44
4	Start	0	—	—	0.00	1639.23	0.00	0.00	1639.23
	Opt.	36	214.40	215.60	0.18	303.00	0.00	1292.69	1595.87

TABLE 3. *Final optimization of the model*

Variable i	Point	n_i	LCL_i	UCL_i	C_I	C_A	C_R	C_S	C_T
1	Start	30	79.25	80.75	$.55	$6344.28	$3146.50	$6016.18	$15,507.51
2		32	47.94	52.06					
3		22	23.70	26.30					
4		36	214.40	215.60					
1	Optimum	36	79.42	80.58	1.13	4127.35	6846.35	3653.21	14,628.04
2		34	48.34	51.66					
3		28	23.98	26.02					
4		214	214.37	215.63					

Existence of a Unique Minimum

In applying a search procedure to the optimization of a mathematical model, one tacitly assumes that the model possesses a single minimum or maximum, as opposed to several local optima. In some cases this condition can be verified analytically. In many cases, however, the existence of a single minium (or maximum) cannot be proven analytically but may be demonstrated by examination of representative example cases. The latter approach was taken in the case of the model presented here. It should be noted that examination of a representative set of example problems does not "prove" that the model possesses a unique minimum (or maximum), but it can indicate that uniqueness of the optimum can be expected in a large number of cases. In this research the authors were not able to identify a case for which the model possessed several local minima.

Two approaches were taken to determine whether local minima existed for several example problems analyzed. First, several starting points were chosen for application of the search procedure. For each example, the optimum indicated by the search for the various starting points clustered in the same neighborhood and resulted in total costs of quality control which were within 0.10% of one another.

The second approach to the analysis of the behavior of the function requires the selection of a random point in the solution space and evaluation of the function on the line segment

between that point and the optimum indicated by the search procedure. This process was repeated for five different randomly selected points for several examples. Specifically, let Δ_i be the maximum deviation from μ_{i0} analyzed for the lower and upper control limits, M_i be the maximum sample size for the ith variable, and the r_{1i}, r_{2i}, and r_{3i} be random numbers associated with the sample size and the lower and upper control limits for the ith variable, and n_i^*, LCL_i^*, UCL_i^* be the associated optimum sample size and lower and upper control limits. Let $\mathbf{X}$ and $\mathbf{X}^*$ be the vectors defined by

$$\mathbf{X} = \begin{Bmatrix} m_1 r_{11} \\ \mu_{10} - \Delta_1 r_{21} \\ \mu_{10} + \Delta_1 r_{31} \\ m_2 r_{12} \\ \mu_{20} - \Delta_2 r_{22} \\ \mu_{20} + \Delta_2 r_{32} \\ \cdot \\ \cdot \\ \cdot \\ m_{k_3} r_{1k_3} \\ \mu_{k_3,0} - \Delta_{k_3} r_{2,k_3} \\ \mu_{k_3,0} + \Delta_{k_3} r_{3,k_3} \end{Bmatrix} \qquad \mathbf{X}^* = \begin{Bmatrix} n_1^* \\ LCL_1^* \\ UCL_1^* \\ n_2^* \\ LCL_2^* \\ UCL_2^* \\ \cdot \\ \cdot \\ \cdot \\ n_{k_3}^* \\ LCL_{k_3}^* \\ UCL_{k_3}^* \end{Bmatrix}$$

Then a convex combination of $\mathbf{X}$ and $\mathbf{X}^*$, $\lambda\mathbf{X} + (1 - \lambda)\,\mathbf{X}^* (0 < \lambda < 1)$, yields a point on the line segment joining $\mathbf{X}$ and $\mathbf{X}^*$.

For each example problem examined the total cost of quality control associated with $\lambda\mathbf{X} + (1 - \lambda)\,\mathbf{X}^*$ proved to be greater than at $\mathbf{X}^*$ indicating that the search technique does in fact identify a point as optimum which is close to the true global optimum. In addition, this analysis showed that

$$C_T[\lambda_1 \mathbf{X} + (1 - \lambda_1)\,\mathbf{X}^*] < C_T[\lambda_2 \mathbf{X} + (1 - \lambda_2)\,\mathbf{X}^*] \tag{19}$$

for $\lambda_1 > \lambda_2$. Equation (19) implies that the cost model probably possesses only one local minimum, at least for the examples treated.

As a corollary to the analysis just described, the authors have analyzed several examples by varying each of the decision variables, one at a time, from their optimal values and evaluating the total cost of quality control as a function of these deviations. This analysis indicates cost is most sensitive to departures of the control limits from optimal for those variables for which inspection is destructive (classes I and II) and least sensitive to departures of sample size from optimal for variables for which inspection is nondestructive (class III). The variation of total expected cost with variation of the decision variables (one at a time) is illustrated in Figures 2 and 3 for the problem in Example 1. For Figure 2, the abscissa is the ratio n_i/n_i^*, $i = 1, 2, 3, 4$. In Figure 3 the abscissa R is the multiple of μ_{i0}, $i = 1, 2, 3, 4$, such that

$$UCL_i = \mu_{i0} + R\mu_{i0} \tag{20}$$

and

$$LCL_i = \mu_{i0} - R\mu_{i0}, \tag{21}$$

and $R = 0.0375$ yields the optimum values of UCL_i and LCL_i.

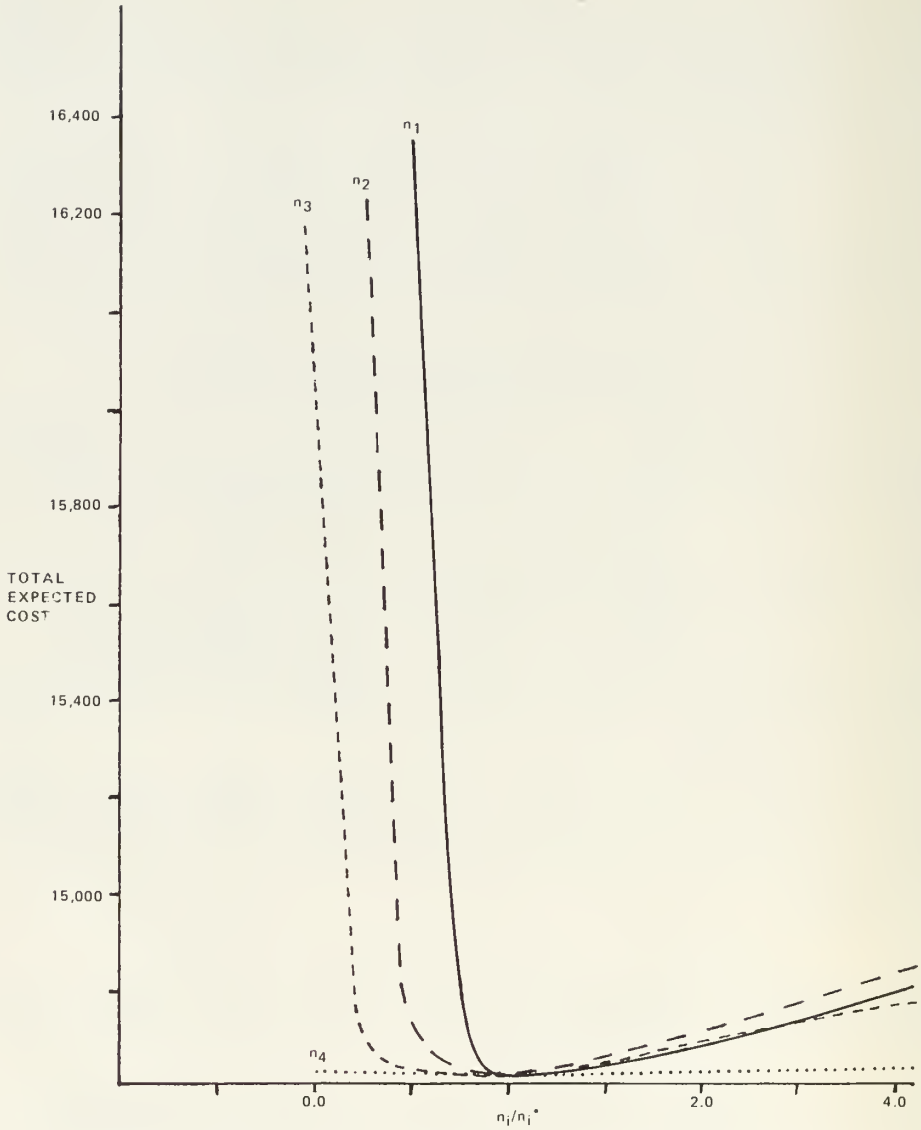

FIGURE 2. Total expected cost of quality control per lot as a function of the sample size n_i, $i = 1, 2, 3, 4$, for deviations from n_i^*, for the problem in Example 1.

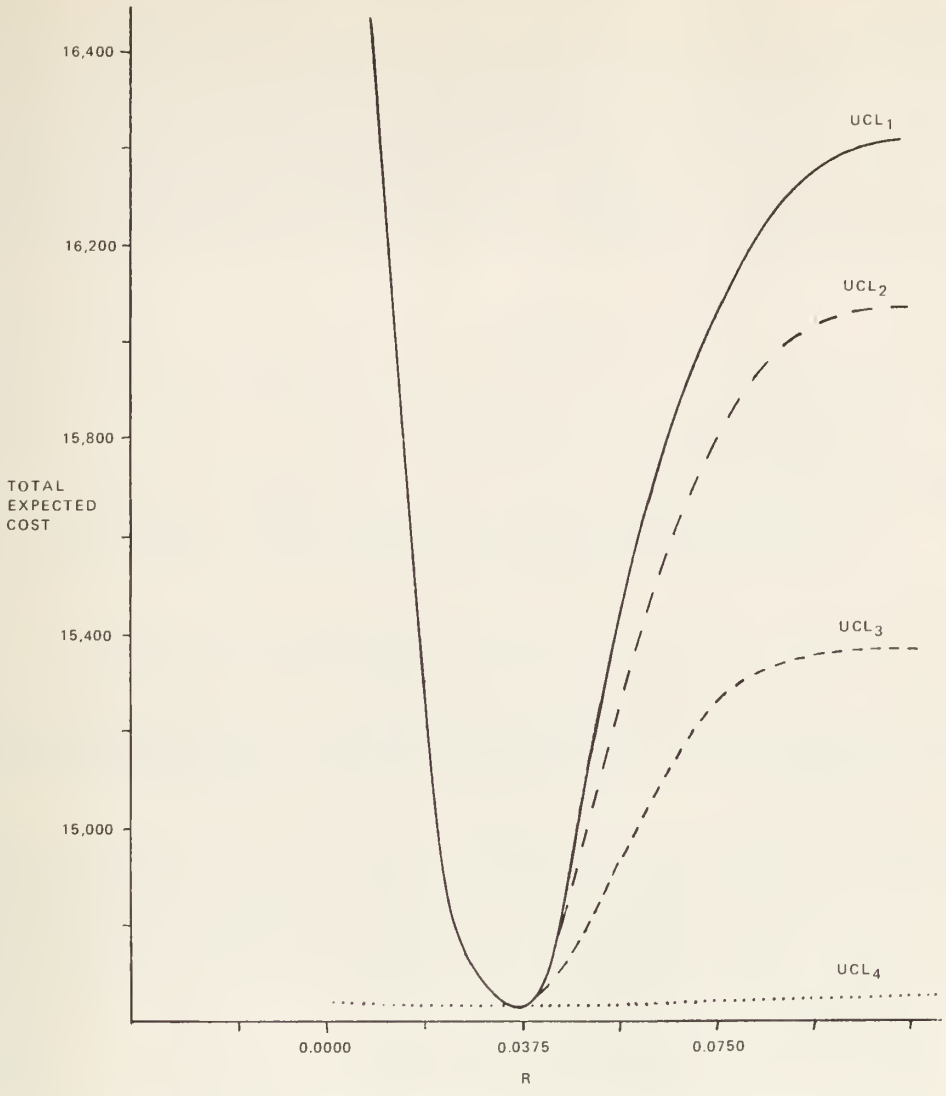

FIGURE 3. Total expected cost of quality control per lot as a function of deviation of UCL_i from UCL_i^* and LCL_i from LCL_i^* for the problem in Example 1.

Sensitivity Analysis

To apply the model presented in this paper, several system parameters must be estimated and probability distributions identified. One of the most difficult estimation problems is that associated with identification of the density function of the lot mean for each of the variables considered. In general, the parameters of these density functions can be estimated through the collection of a moderate volume of data. However, identification of the shape of each density function may require an extensive and expensive data-collection effort.

The research of Latimer, Bennett, and Schmidt [18] and Schmidt and Bennett [29] has indicated that in many cases the optimal sampling plan and the total cost of quality control are not sensitive to the shape of the distribution of the random variables which are subjected to

control. However, a method for analyzing the sensitivity of the model to distributional shape is necessary if this analysis is to be carried out in specific applications of the model. The result of such an analysis will indicate the extent of the data-collection effort required. That is, if the model proves to be insensitive to distributional shape for a given application, then the analyst will require only a rough idea of the shape of the density functions involved. On the other hand, if the model proves to be sensitive to distributional shape, then an extensive data collection effort may be in order to specify precisely the density function of the lot mean for each variable to be controlled.

If we let

$f(\mu_i)$ = p.d.f. of μ_i,
C_T^* = total expected cost of quality control per lot under the optimum set of sampling plans if $f(\mu_i)$ is known,
$\hat{f}(\mu_i)$ = assumed p.d.f. of μ_i,
$\hat{C}_T^*$ = total expected cost of quality control per lot under the optimum set of sampling plans *derived* when $\hat{f}(\mu_i)$ is assumed, but *evaluated* using $f(\mu_i)$,

then

$$P = \frac{\hat{C}_I^* - C_T^*}{C_T} (100) \tag{22}$$

is the percent loss realized as a result of using the assumed density function $\hat{f}(\mu_i)$ instead of the actual density function $f(\mu_i)$. To illustrate the sensitivity analysis suggested here, consider the following example.

EXAMPLE 2: The problem considered here is the same as that presented in Example 1. The assumed density function of μ_i, $\hat{f}(\mu_i)$, is normal, with mean and variance given in Table 1, $i = 1, 2, 3, 4$; it is assumed that the mean and variance have been estimated with reasonable precision. Now we choose a density function $f(\mu_i)$ having the same mean and variance as $\hat{f}(\mu_i)$ but a different shape for all i. For this example we will choose distributions having the shapes shown in Figure 4. That is, C_T^* is evaluated when $f(\mu_i)$, $i = 1, 2, 3, 4$, has the shape given in Figure 4b. Then $\hat{C}_T^*$ is evaluated with this same set of density functions, $f(\mu_i)$, $i = 1, 2, 3, 4$. The percent loss realized is then calculated by equation (22). The process is then repeated with $f(\mu_i)$, $i = 1, 2, 3, 4$, taking on the shapes given in Figures 4c, 4d, and 4e successively. Since the distributional shapes given in Figures 4b, c, d, and e represent rather signficant departures from the assumed shape given in Figure 4a, calculation of the percent loss in each case should indicate the sensitivity of the model for the application given in Example 1 to significant errors in identifying the true shape of $f(\mu_i)$, $i = 1, 2, 3, 4$. The results of this analysis are summarized in Tables 4 and 5. Table 4 presents the optimum cost of quality control per lot when $f(\mu_i)$ is known for each of the distributions shown in Figures 4b, c, d, and e. Table 5 gives the percent loss due to determination of the optimum sampling plan when the density functions of μ_i are assumed to be normal, $i = 1, 2, 3, 4$, but actually assume the shapes given in Figures 4b, c, d, and e.

The results summarized in Tables 4 and 5 indicate that errors in identifying the shape of the distribution of the lot mean are not serious, at least for this example. While general conclusions cannot be drawn regarding the sensitivity of the model and the optimum soluion to distributional shape, these results are not surprising in light of similar results given by Schmidt and Bennett [29] and Latimer, Bennett, and Schmidt [18]. In any case, the method presented for sensitivity analysis may be applied to any specific case where the model is used.

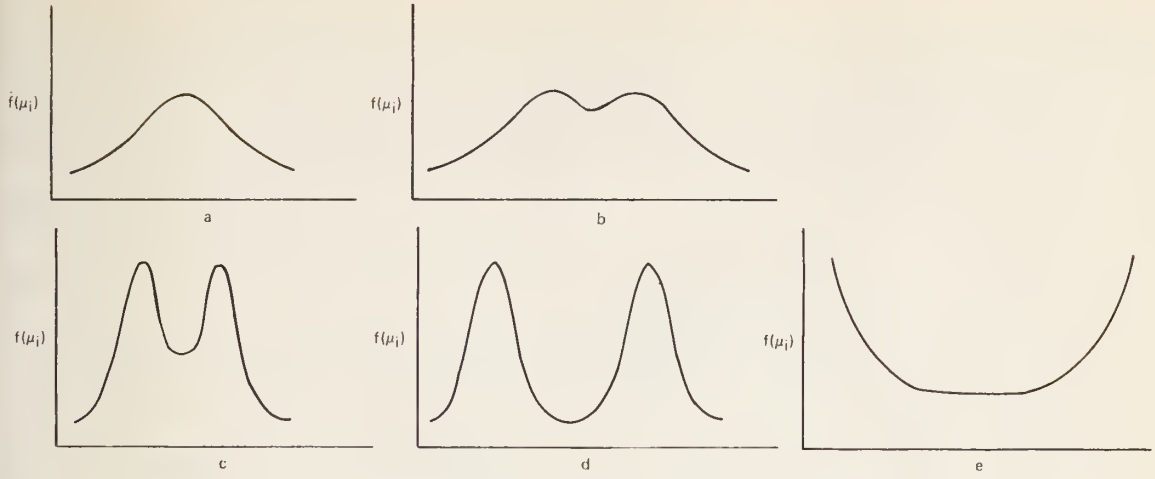

FIGURE 4. Assumed (a) and true (b,c,d,e) density functions used for the sensitivity analysis in Example 2.

TABLE 4. *Optimum sampling plans and expected total cost of quality control per lot for the distributions shown in Figures 4b, c, d, and e*

$f(\mu_i)$	Variable Number	Sample Size	Control Limits (Lower, Upper)	Total Expected Cost
b	1	38	(19.393, 20.606)	$14,837.43
	2	38	(48.266, 51.733)	
	3	38	(3.978, 6.021)	
	4	105	(14.224, 15.784)	
c	1	48	(19.410, 20.591)	$15,358.03
	2	46	(48.303, 51.695)	
	3	42	(4.026, 5.974)	
	4	136	(14.312, 15.694)	
d	1	0	—	$16,061.91
	2	0	—	
	3	0	—	
	4	234	(14.433, 15.554)	
e	1	0	—	$15,739.88
	2	0	—	
	3	0	—	
	4	326	(14,298,15.700)	

TABLE 5. *Percent loss due to assuming that μ_i, $i = 1, 2, 3, 4$, is normally distributed when μ_i has the density function $f(\mu_i)$, $i = 1, 2, 3, 4$, for each of the shapes given in Figures 4b, c, d, and e*

$f(\mu_i)$	C_T^*	$\hat{C}_T^*$	P
b	$14,837.43	$14,872.21	0.23%
c	15,358.03	15,402.16	0.29%
d	16,0.61.91	16,261.81	1.24%
e	15,739.88	16,316.37	3.66%

BIBLIOGRAPHY

[1] ASQC Standard A3 — 1964, "Glossary of General Terms Used in Quality Control," (American Society for Quality Control, 1964).

[2] Bather, J. A., "Control Charts and the Minimization of Costs," Journal of the Royal Statistical Society, Series B, *25* (1963).

[3] Bowker, A. H., and H. P. Goode, *Sampling Inspection by Variables*, (McGraw-Hill, New York, 1952).

[4] Dodge, H. F., and H. G. Romig, "A Method of Sampling Inspection," Bell System Technical Journal *8*, 613-631, (October 1929).

[5] Duncan, A. J., "The Economic Design of $\bar{X}$ Charts Used to Maintain Current Control of a Process," Journal of the American Statistical Association *51*, 228-242 (June 1956).

[6] Duncan, A. J., *Quality Control and Industrial Statistics*, (Richard D. Irwin, Inc., Homewood, Illinois, 1965).

[7] Duncan, A. J., "The Economic Design of $\bar{X}$ Charts When There is a Multiplicity of Assignable Causes," Journal of the American Statistical Association *66*, 107-121 (March 1971).

[8] Ervin, N. E., "Simplified Variables Sampling—Graphical Method," ASQC Conference Transactions, 767-776 (American Society for Quality Control, 1966).

[9] Ghare, P. M., and P. E. Torgersen, "The Multicharacteristic Control Chart," Journal of Industrial Engineering *19*, 269-272 (June 1968).

[10] Gibra, Issac N., "Economically Optimal Determination of the Parameters of $\bar{X}$ Control Charts," Management Science *17*, 635-646 (May 1971).

[11] Goel, A. L., et al., "An Algorithm for the Determination of the Economic Design of $\bar{X}$ Charts Based on Duncan's Model," Journal of the American Statistical Association *63*, 304-320 (March 1968).

[12] Hald, A., "The Compound Hypogeometric Distribution and a System of Single Sampling Inspection Plans Based on Prior Distributions and Costs," Technometrics *2*, 275-339 (1960).

[13] Hamaker, H. C., "Economic Principles in Industrial Sampling Problems — A General Introduction," Bulletin of the International Statistical Institute *33*, 105-122, (1951).

[14] Hill, I. D., "The Economic Incentive Provided by Sampling Inspection," Applied Statistics *9*, 69-81 (1960).

[15] Horsnell, G., "Economical Acceptance Sampling Schemes," Journal of the Royal Statistical Society, Series A *120*, 148-201 (1957).

[16] Hotelling, H., "Multivariate Quality Control," in *Techniques of Statistical Analysis*, C. Eisenhart, et al., eds. (McGraw-Hill, New York 1947).

[17] Knappenberger, J. A., and A. H. E. Grandage, "Minimum Cost Quality Control Tests," AIIE Transactions *1* (1969).

[18] Latimer, B. A., G. K. Bennett, and J. W. Schmidt, "The Economic Design of a Dual Purpose Multicharacteristic Quality Control System" AIIE Transactions *5* (3) (1973).

[19] Lieberman, G. J., and G. J. Resnikoff, "Sampling Plans for Inspection by Variables," Journal of the American Statistical Association *50*, 457-516 (1955).

[20] Montgomery, D. C., and P. J. Klatt, "Minimum Cost Multivariate Quality Control Tests AIIE Transactions *4*, 103-110 (1972).

[21] Mood, A. M., "On the Dependence of Sampling Inspection Plans Based Upon Population Distributions," Annals of Mathematical Statistics *14*, 415-425 (1943).

[22] Owen, D. B., "Summary of Recent Work on Variables Acceptance Sampling With Emphasis on Non-Normality," Technometrics, *11*, 631-637 (1969).

[23] Owen, D. B., "Variables Sampling Plans Based on the Normal Distribution," Technometrics *9*, 417-423 (1967).

[24] Owen, D. B., "One-Sided Variables Sampling Plans," Industrial Quality Control *22* 450-456 (1966).

[25] Palazzi, A., "Statistical Inspection by Variables — Some General Observations," Quality, EOQC Journal *15* 6-7 (Spring 1971).

[26] Pfanzagl, J., "Sampling Procedures Based on Prior Distributions and Costs," Technometrics *5*, 47-61 (1963).

[27] Pfanzagl, J., and W. Schuler, "The Efficiency of Sequential Sampling Plans Based on Prior Distributions and Costs," Technometrics *12*, 299-310 (1970).

[28] Quality Cost — Cost Effectiveness Technical Committee, ASQC, "Quality Costs — What and How," (American Society for Quality Control, 1967).

[29] Schmidt, J. W., and G. K. Bennett, "A Cost Model for the Simultaneous Control of Several Attributes," AIIE Transactions 4, 194-199 (1972).

[30] Schmidt, J. W., and R. E. Taylor, "A Dual Purpose Cost Based Quality Control System," Technometrics 15, 151-166 (1973).

[31] Shewhart, W. A., *Economic Control of Quality of Manufactured Product*, (Van Nostrand, New York, 1931).

[32] Schilling, E. G., "Variables Sampling and Military Standard 414," ASQC Conference Transactions, 175-188 (1970).

[33] Shuler, W., "Multistage Sampling Procedures Based on Prior Distributions and Costs," Annals of Mathematical Statistics *38*, 464-470 (1967).

[34] Shakun, M. F., "Multivariate Acceptance Sampling Procedures for General Specification Ellipsoids," Journal of the American Statistical Association *60*, 905-913 (1965).

[35] Sittig, J., "The Economic Choice of Sampling Systems in Acceptance Sampling," Bulletin of the International Statistical Institute *33*, 51-84 (1951).

[36] Smith, B. E., "The Economics of Sampling Inspection," Industrial Quality Control *21* 453-458 (1965).

[37] Srivastava, A. B. L., "Variables Sampling Inspection for Non-Normal Samples," Journal of Scientific and Engineering Research *5* 145-152, (1961).

[38] Stacy, E. W., et al., "Determining Economic Sampling Plans," IBM Systems Journal *8*, 220-244 (1969).

[39] Weibull, W., "A Method of Determining Inspection Plans on an Economic Basis," Bulletin of the International Statistical Institute *33*, 85-104 (1931).

[40] Weiler, H., "On the Most Economical Sample Size for Controlling the Mean of a Population," Annals of Mathematical Statistics *23*, 247-254 (1952).

[41] Wetherill, G. B., *Sampling Inspection and Quality Control*, (Methuen, London, 1969).

[42] Wilde, D. J., and C. S. Beightler, *Foundations of Optimization*, (Prentice-Hall, Englewood Cliffs, N. J., 1967).

MARKOV CHAIN WITH TABOO STATES AND RELIABILITY

G.L. Sriwastav

Department of Statistics
Dibrugarh University
Dibrugarh-786004, India

S.N.N. Pandit

Department of Mathematics,
Regional Engineering College
Warangal-506004, India

ABSTRACT

A computationally feasible matrix method is presented to find the first-passage probabilities in a Markov chain where a set of states is taboo during transit. This concept has been used to evaluate the reliability of a system whose changes in strength can be thought of as a Markov chain, while the environment in which it is functioning generates stresses which can also be envisaged as another Markov chain.

1. INTRODUCTION

The "strength" $X(t)$ of a system can be thought of as a random process, and the "stresses" $Y(t)$ acting on the system can be thought of as another stochastic process. The system will work as long as $X(t) \geqslant Y(t)$, and will fail at t^* defined by

$$X(t) \geqslant Y(t),\ t < t^* \text{ and } X(t^*) < Y(t^*)$$

The reliability problem we are proposing here is: Find the probability distribution of t^*, the time to failure of the system, given an adequate description of the stochastic processes $X(t)$ and $Y(t)$.

The problem can also be thought of as the time to first zero-crossing of the difference process $Z(t) = X(t) - Y(t)$, $Z(t)$ being positive. The problem considered in the present study is a discrete version of the above problem, and can be described as follows: We replace the continuous stochastic processes $X(t)$ and $Y(t)$ by discrete Markov chains $X(r)$ and $Y(r)$, $r = 0, 1, 2, \ldots$, with stationary transition probabilities, where $X(r)$ and $Y(r)$ take values from discrete sets.

We assume that, at each step $r = 0, 1, 2, \ldots$, the system will have one of the possible strength levels identified by the numbers $1, 2, 3, \ldots, n$. The system strength $X(r)$ transits from one step to the next as a Markov chain with stationary transition probabilities matrix $P = (p_{ij})$, $i, j = 1, 2, \ldots, n$.

Similarly, at each r, the "environment" impinges a stress $Y(r)$, identified by the numbers $1, 2, \ldots, m$, which is again a Markov chain with stationary transition probabilities matrix $Q = (q_{lk})$, $l, k = 1, 2, \ldots, m$.

Let us define the "compound state" of the system as the pair (α, β), where α is the strength state and β is the stress state; it gives the state of the system at any step.

Let $Z = \{z(i,l; j,k)\} = P \times Q$, the Kronecker product of P and Q. It is a square matrix of order $N = n \times m$ and gives the one-step transition probabilities of transition from the compound state (i,l) to the compound state (j,k).

The compound states of the system are classified into "survival" set S and "failure" set F as follows:

$$(i, l) \in S \text{ if } i \geqslant l \text{ and } (i, l) \in F \text{ if } i < l.$$

The problem is to find the system-survival function, i.e., the chance that the system does not enter F set in the first r steps, $r = 0, 1, 2, \ldots, \infty$, conditional to a specified initial state.

The question essentially is one of determining the first-entry probabilities with respect to Z, taking the failure states as taboo states.

The first-entry problem with a prescribed taboo-state set is, of course, not new. For instance, the well known relation [2]

$$f_{\alpha\beta}^{(r+1)} = p_{\alpha\beta}^{r+1} - \sum_{k=1}^{r} f_{\alpha\beta}^{(r+1-k)} \, p_{\beta\beta}^{k},$$

giving the probability for the first entry to state β at the $(r+1)$th step, gives the answer to our problem when the taboo set is just the single state β. However, it is obvious that the above formula cannot be used when the taboo set consists of more than one state.

Chung [1] has given a recurrence relation

$$p_{\alpha\beta}(r, F) = p_{\alpha\beta}(r, k \cup F) + \sum_{\nu=1}^{r-1} p_{\alpha\beta}(\nu, k \cup F) \, p_{k\beta}(r-\nu, F)$$

where F is the taboo set and $k \notin F$.

However, this relation is essentially conceptual and is of no help at all in the actual determination of failure probabilities; it is obvious that the computation of $p_{\alpha\beta}(r, F)$ by this method requires a knowledge of $p_{\alpha\beta}(r, k \cup F)$ as well as of $p_{\alpha\beta}(\nu, k \cup F)$ and $p_{k\beta}(r-\nu, F)$; thus begging the question itself. Hence the need for a constructive procedure to solve the problem.

2. FIRST-PASSAGE PROBABILITIES IN THE PRESENCE OF A TABOO SET

We shall now present a matrix approach to solve the problem. The relevant concepts are introduced first and then the procedure is developed. It is then illustrated by a simple numerical example. Let Z be a transition probability matrix of order N and let $F = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ be the set consisting of k states, among N states, which are taboo in transition.

We define a deficient matrix Z' relative to the taboo set F. Z' is obtained from $Z = \{z(i,l;j,k)\}$ by replacing all $z(i,l;j,k)$ by zeros when $i < l$ or $j < k$ or both; only those $z(i,l;j,k)$ when $i \geqslant l$ and $j \geqslant k$ will remain nonzero in Z'.

Then, obviously, $p_{ij}(r+1,F)$, the probability of the system initially at i reaching state j at the $(r+1)$th step, avoiding visits to states in F during the first r steps, is given by $p_{ij}(r+1, F) = (Z')^r Z$.

Now Z' is a nonnegative matrix with each of its row totals < 1, since by assumption P and Q are positive matrices. Hence , by an extension of Frobenius' theorem [3], all characteristic roots of Z' are necessarily less than 1 in magnitude. Consequently

$$\lim_{r\to\infty} (Z')^r = 0 \tag{2.1}$$

and $(I-Z')$ is nonsingular.

Since $(Z')^r$ gives the probabilities of r step transitions from a nontaboo state to a nontaboo state, avoiding entry into taboo states during transit, we have $(Z')^r Z$ giving transition probabilities of reaching from a nontaboo state α to any state β in $r+1$ steps, avoiding taboo states in the first r steps.

Let $U(S)$ be a column vector of order N having 1 in places corresponding to nontaboo states S and zero elsewhere, and let $U(F)$ be another column vector of order N with 1 in places corresponding to taboo states F and zero elsewhere. Then $(Z')^{r-1} Z\, U(S)$ gives the probabilities that starting from state α, $\alpha = 1, 2, \ldots, N$, the system does not fail in the first r steps. Similarly, $(Z')^{r-1} Z\, U(F)$ gives probabilities that starting from state α failure occurs for the first time at the rth step.

As $\lim_{r\to\infty} (Z')^r = 0$, the system must fail ultimately. Hence, the event of nonfailure in the first r steps is identical with the event of (first) failure in one of the steps $r+1, r+2, \ldots, \infty$. Thus we have,

$$(Z')^{r-1} Z\, U(S) = \sum_{\nu=0}^{\infty} (Z')^{r+\nu} Z\, U(F)$$

Since,

$$Z\, U(S) = Z'\, U(S)$$

$$(Z')^{r-1} Z' U(S) = (Z')^r \sum_{\nu=0}^{\infty} (Z')^{\nu} Z\, U(F),$$

i.e.,

$$(Z')^r U(S) = (Z')^r (I-Z')^{-1} Z\, U(F) \tag{2.2}$$

We remark that for computational purposes we can delete the zero rows and columns of Z' with corresponding changes in the other matrices. We see that the deletion of the corresponding rows from the column matrix $Z\, U(F)$, along with Z', will not change the right-hand side of equation (2.2). Let $U_a(S)$ be the abridged form of $U(S)$ obtained by deleting zero rows in $U(S)$. Let Z'_a be the abridged form of Z'. Then Z'_a is nonsingular, and so is $(I-Z'_a)$, as Z'_a has all its characteristic roots less than 1 in magnitude. Hence from equation (2.2) we have

$$U_a(S) = (I - Z'_a)^{-1} Z_a U_a(F) \tag{2.3}$$

In equation (2.2) r is a random variable representing the number of steps up to which the system does not fail. Its conditional probability distribution on condition that the system starts in state α ($\alpha = 1, 2, \ldots, N$), is given by the row corresponding to state α of the column vector sequence

$$(Z'_a)^r Z_a U_a(F), \quad r = 0, 1, 2, \ldots, \infty.$$

Hence, for a given r and the Z matrix, conditional reliability of the system can be evaluated from (2.2) on condition that the system starts in state α. Unconditional reliability $R(r)$ of the system for r steps is then given by $R(r) = p_{oa} (Z'_a)^r U_a(S)$, where p_o is the row matrix of order N giving the probability distribution of initial states of the system, and p_{oa} is its abridged form.

The moment generating function of r is given by

$$M(\theta) = \sum_{r=0}^{\infty} e^{r\theta} p_{oa} (Z'_a)^r Z_a U_a(F) = p_{oa} (I - e^{\theta} Z'_a)^{-1} Z_a U_a(F).$$

The conditional νth factorial moment of r can be directly obtained as

$$\begin{aligned} \mu_{[\nu]} &= E[r(r-1) \ldots (r - \nu + 1) / \alpha] \\ &= (Z'_a)^r (I - Z'_a)^{-\nu-1} Z_a U_a(F) \\ &= (Z'_a) (I - Z'_a)^{-\nu} U_a(S), \text{ by (2.3).} \end{aligned}$$

The νth unconditional factorial moment is

$$\mu'_{[\nu]} = p_{oa} (Z'_a) (I - Z'_a)^{-\nu} U_a(S).$$

An Illustration

Let us consider an example with strength and stress transition probabilities matrices P and Q, respectively, as follows:

$$P = (p_{ij}) = \begin{bmatrix} 0.3 & 0.7 \\ 0.2 & 0.8 \end{bmatrix} \text{ and } Q = (q_{lk}) = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.6 & 0.3 \end{bmatrix}.$$

Then,

$$Z = P \times Q$$

States

	1	2	3	4	5	6
1	0.18	0.09	0.03	0.42	0.21	0.07
2	0.12	0.12	0.06	0.28	0.28	0.14
= 3	0.03	0.18	0.09	0.07	0.42	0.21
4	0.12	0.06	0.02	0.48	0.24	0.08
5	0.08	0.08	0.04	0.32	0.32	0.16
6	0.02	0.12	0.06	0.08	0.48	0.24

Clearly states 2, 3, and 6 are failure states. Hence the abridged deficient matrix is,

$$Z'_a = \begin{bmatrix} 0.18 & 0.42 & 0.21 \\ 0.12 & 0.48 & 0.24 \\ 0.08 & 0.32 & 0.32 \end{bmatrix}$$

and

$$(I - Z'_a)^{-1} = \begin{bmatrix} 0.2768 & 0.3528 & 0.2100 \\ 0.1008 & 0.5408 & 0.2220 \\ 0.0800 & 0.2960 & 0.3760 \end{bmatrix} \div 0.167840.$$

$\bar{U}(S) = (1\ 0\ 0\ 1\ 1\ 0)$, and $\bar{U}(F) = (0\ 1\ 1\ 0\ 0\ 1)$, where $\bar{U}(S)$ and $\bar{U}(F)$ are transposes of $U(S)$ and $U(F)$ respectively.

It can be easily verified that

$$(I - Z'_a)^{-1}\ Z_a\ U_a(F) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

where rows corresponding to the failure states are deleted.

We may note that to calculate the probability of success in the first r steps it is not necessary to calculate the powers of Z'_a. The following recurrence relationship can be used:

Let $(Z'_a)^r\ U_a(S) = q(r,S)$

=the reliability vector at the r*th step,*

then $q(r+1,\ S) = Z'_a\ q(r,\ S)$

=the reliability vector at the $(r+1)$*th step.*

Obviously $q(0,\ S)$ = transpose of $(1,\ 1,\ 1)$.

For the given example, for $r = 0,\ 1,\ 2,\ 3,\ 4$, we have

Z'_a			q(r, s)				
			$r = 0$	$r = 1$	$r = 2$	$r = 3$	$r = 4$
0.18	0.42	0.21	1	0.81	0.6490	0.518148	0.412641
0.12	0.48	0.24	1	0.84	0.6732	0.536472	0.427178
0.08	0.32	0.32	1	0.72	0.5640	0.447885	0.356447

Let the initial-state probabilities of the system be given by $p_o = (0.3,\ 0.1,\ 0.2,\ 0.1,\ 0.2,\ 0.1)$, and the abridged form by $p_{oa} = (0.3,\ 0.1,\ 0.2)$.

Then $R(r)$, reliability at the rth step, is given by

$$R(1) = 0.471,\ R(2) = 0.375060,\ R(3) = 0.298669,\ R(4) = 0.237799.$$

The conditional expectation of r on the condition that the system starts in state α is given by

$$E(r/\alpha) = Z'_a\,(I-Z'_a)^{-1}\,U_a(S) = \begin{bmatrix} 4.002383 \\ 4.145377 \\ 3.480458 \end{bmatrix}.$$

Hence the unconditional expectation of r is given by

$$E(r) = 1.2007149 + 0.4145377 + 0.6960916 = 2.3113442.$$

Similarly, $E(r^2) = 20.3167038.$

Hence $V(r) = 14.9743918,$

and $S.D.\,(r) = 3.8696759.$

REFERENCES

[1] Chung, K.L., *Markov-Chains With Stationary Transition Probabilities* (Springer-Verlag, Berlin, 1960).

[2] Feller, W., *An Introduction to Probability Theory and Its Application*, Vol. I, 2nd ed. (Wiley, New York, 1957).

[3] Gantmacher, F.R., *The Theory of Matrices*, Vol. I (Chelsea Publishing Company, New York, 1959).

OPTIMAL SAMPLING AND THE RIGHT-HAND SIDE OF A LINEAR PROGRAM: SENSITIVITY ANALYSIS REVISITED

Andrew F. Daughety

Graduate School of Management
Northwestern University
Evanston, Illinois

ABSTRACT

This paper examines right-hand side sensitivity analysis in linear programming as a problem in optimal sampling. Specifically, the insensitivity point of a solution is defined as the point at which the expected gain from increased accuracy in the prediction of a resource level is equal to the expected cost of procuring the information. The problem is structured using the rudiments of optimal statistical decision theory.

INTRODUCTION

This paper is concerned with sensitivity analysis in linear programs. Usually, when we have considered a solution's sensitivity, we have found (for example) the range that the right-hand side could vary before a change in basis was required. The implicit motivation was that if the right-hand side changed, and the range of values over which the optimal basis was applicable was in some sense "small," then we would be operating with a solution to the wrong problem. In this paper we will attempt to formalize this notion.

We will argue that the real problem is to find the value for the right-hand side that best trades off potential losses (for using an incorrect value) with the cost of obtaining the improved solution. We will limit ourselves to changing only one element of the right-hand side, which will be a sample mean instead of the unknown, true value.

Two problems will be discussed. In the first we assume some data has been used to form an estimate and we formulate the function to be minimized that indicates how much more data should be gathered so as to "desensitize" the optimal solution. We will then examine the problem of finding the overall sample size in order to produce an insensitive solution.

THE BASIC PROBLEM

Consider the following linear program:

$$\text{(LP)} \qquad \max \quad px$$
$$\text{subject to} \quad Ax \leqslant b,$$
$$x \geqslant 0,$$

where p and x are k-vectors, b is a nonnegative m-vector and A is a $m \times k$ matrix. We will assume that p and A are perfectly known, as are all elements of b except for b_i, which will have to be estimated. Finally, we assume that n data points are used to construct an unbiased estimate of b_i, which we shall symbolize as $\hat{b}_i(n)$. While we have in mind the sample mean as the estimator, it will be seen that other estimators (such as regression coefficients) could also be used. However, for simplicity we take

$$\hat{b}_i(n) \sim N\left(b_i, \frac{\sigma^2}{n}\right) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left[\frac{-n}{2}\left(\frac{\hat{b}_i(n) - b_i}{\sigma}\right)^2\right],$$

where b_i is the true value of the amount of the i^{th} resource on hand. Let the optimal solution to (LP) when $\hat{b}_i(n)$ is used† for b_i be $\hat{x}$ with associated basis matrix $\hat{B}$. Further let the optimal solution to (LP) when the true value of the i^{th} resource (namely b_i) is used be x^* with associated basis matrix B^*.

In general, our intuitive feeling is that a solution is sensitive to the right-hand side value if "small" changes in the value might require a change of basis. We will assume that the changes in the value of $\hat{b}_i(n)$ came about from changes in the data used to estimate b_i as n is changed.

Figure 1 illustrates the problem to be discussed. The solid lines represent constraints while the dashed line represents the objective function. Two bell-shaped curves are drawn, centered about the *true value* of the right-hand side of the third constraint. If this value were known then the optimal solution would occur at point A. Let the curve labelled 1 represent the density function of the sample mean given that n data points have been used to form the sample. For this right-hand side the solution is sensitive in the sense that the sample mean might easily place the constraint considerably to the left or right of the true location, e.g., it might end up at B or C.

There are basically two approaches to the problem. One method would be essentially to chance-constrain the program. This would include increasing the sample size until the estimated variance of the sample mean were in some sense "small", i.e., the probability of a sample mean occurring outside of some interval about the true value would be less than or equal to α. This is represented in Figure 1 by the curve labelled 2. This requires the setting of α at some arbitrary level and thus really does not clarify the sensitivity problem at all. Now the solution is sensitive to α, which was arbitrary.

A second approach is available if the cost of gathering data is expressible in the same units as the objective function, e.g., dollars. We will examine this approach in detail in the rest of the paper. We will assume that a data cost function can be obtained and that it faithfully reflects the costs of acquiring, processing, storing and providing sample-mean estimates. We assume that costs are a function of the amount of data only and thus $c(n)$ is the cost for providing $\hat{b}_i(n)$.

In what follows we will develop loss functions and state the sensitivity problem as finding the sample size that minimizes expected loss plus cost. Thus, rather than setting an arbitrary parameter, we will consider a solution to be insensitive when the marginal expected gain from extra information is just balanced by its marginal cost.

† We refer to the whole right-hand side in this case as $\hat{b}(n)$.

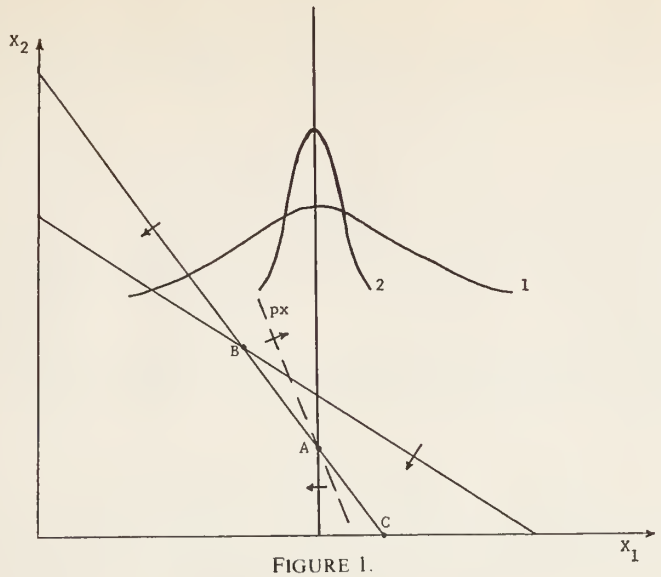

FIGURE 1.

CONSTRUCTING AN *Ex Ante* LOSS FUNCTION

Consider a decision maker who wishes to solve (LP) with $\hat{b}(n)$ as the right-hand side. He may choose to produce the activities associated with $\hat{B}$ on the assumption that $\hat{b}(n)$ is available. In this case he will either not use $b - \hat{b}(n)$ resources or he will run out early (when $b < \hat{b}(n)$).

Let b be the true vector of resource levels with i^{th} element b_i and $\hat{b}$ be the vector with $\hat{b}_j = b_j$ $j \neq i$ and $\hat{b}_i = \hat{b}_i(n)$. We define the loss associated with using $\hat{b}$ instead of b as $L(b,\hat{b})$ and will take it to be:

$$L(b,\hat{b}) = \begin{cases} \{\max px \mid Ax \leqslant b,\ x \geqslant 0\} & \\ \quad - \{\max px \mid Ax \leqslant \hat{b},\ x \geqslant 0\}, & \text{if } b \geqslant \hat{b}, \\ \{\max px \mid Ax \leqslant b,\ x \geqslant 0\} & \\ \quad - \{\max px \mid Ax \leqslant b,\ x_{\hat{B}} \geqslant 0,\ x_{N\hat{B}} = 0\}, & \text{if } b \leqslant \hat{b}. \end{cases}$$

where $(x_{\hat{B}} \geqslant 0,\ x_{N\hat{B}} = 0)$ means that only those vectors represented by $\hat{B}$ are free to be nonnegative. The loss function indicates that when $b \geqslant \hat{b}$ the *ex ante* loss will be taken to be the difference between the profit that would have been obtained if it were known that the resource were actually b and the profit that is expected to be obtained by using the program associated with $\hat{B}$ and resource level $\hat{b}$. The left over resources will be considered waste and will be charged to the solution as loss. If $b < \hat{b}$ then we have overestimated the amount of resource. The *ex ante* loss that will be incurred is the difference in profit between what would have been obtained if the resource had been properly estimated and what is expected to be obtained given that we will use the basis matrix $\hat{B}$ but have only b in actual resources.

Let π be the vector of simplex multipliers associated with the optimal feasible basis $(\cdot)$. Then the loss function can be written as:

$$L(b,\hat{b}) = \begin{cases} \pi_{B^*} b - \pi_{\hat{B}} \hat{b}, & b \geqslant \hat{b}, \\ \pi_{B^*} b - \pi_{\hat{B}} b, & b \leqslant \hat{b}. \end{cases}$$

This may be further rewritten as

$$(1) \qquad L(b,\hat{b}) = \begin{cases} (\pi_{B^*} - \pi_{\hat{B}})\, b + (b - \hat{b})\, \pi_{\hat{B}}, & b \geqslant \hat{b}, \\ (\pi_{B^*} - \pi_{\hat{B}})\, b, & b \leqslant \hat{b}. \end{cases}$$

Thus, the loss function can be viewed as being composed of the loss due to using the wrong basis plus the loss due to overestimation of resource availability.

We of course do not know b. We therefore assume that the decision maker or analyst has some prior information about b_i, namely that $b_i \sim N(\bar{b}, \bar{\sigma}^2)$. Thus, the posterior density function on b_i given $\hat{b}_i$, is:

$$(2) \qquad f[b_i|\hat{b}_i(n)] = N([n\hat{b}_i(n)\bar{\sigma}^2 + \bar{b}\sigma^2] \,/\, (n\bar{\sigma}^2 + \sigma^2), \bar{\sigma}^2\sigma^2/(n\bar{\sigma}^2 + \sigma^2)).$$

With the above in mind we are in a position to make the general statement in the introduction more concrete. We will consider a solution to (LP) to be insensitive relative to $\hat{b}_i(n+s)$ if the expected loss associated with $\hat{b}_i(n+s)$ plus cost of sampling s data points $c(s)$ is minimal. Thus, we wish to choose s such that we minimize loss plus cost, i.e.,

$$(3) \qquad \min_s \int L(b,\hat{b}(n+s))\, f[b_i|\hat{b}_i\ (n+s)]db + c(s).$$

The solution to (3), s^*, would be the amount of data to add as to make the optimal solution to (LP) insensitive. However, in that we do not know how $\hat{b}_i(n+s)$ will change with s, we must use our best estimate of the mean, namely the mean of the posterior $E[b|\hat{b}(n)]$. Thus we will find s^* so as to minimize

$$(4) \qquad \int \cdot L(b,E[b|\hat{b}(n)])\, f_s\ [b_i|\hat{b}_i(n)]db + c(s),$$

where

$$f_s\ [b_i|\hat{b}_i(n)] = N(E[b_i|\hat{b}_i(n)],\ \bar{\sigma}^2\ \sigma^2\ /\ [(n+s)\,\bar{\sigma}^2\ + \sigma^2])$$

Minimizing (4) is a heuristic approach to solving (3). In (4) we have used the best estimate of the mean and are only altering the variance of the posterior density function. Using a parametric programming routine on the linear program optimization package will yield the simplex multipliers for all values of b_i. Thus, as is typical in these formulations (see [2]), one would compute the value of (4) for different values of s and select that s that yields at least a local minimum.

A MORE GENERAL PROBLEM

The above procedure raises a significant point, namely the fact that it assumes that n data points have *already* been sampled and used to form $\hat{b}_i(n)$. We would in fact like to find the total sample size in a one-shot approach, thereby avoiding the possibility of having oversampled to begin with.

We will assume here that both parameters of the independent normal process (i.e., b_i and σ_i^2) are unknown and must be treated as random variables. A sample of n data points $(b_{i1}, \ldots, b_{in})$ will be taken, from which the following statistics could be computed:

$$m \equiv \frac{1}{n} \sum_j b_{ij}$$

$$\mathrm{v} \equiv \frac{1}{n-1} \sum_j (b_{ij} - m)^2$$

$$\nu \equiv n - 1$$

Thus, we wish to find n that again minimizes expected loss plus cost $c(n)$. To formulate the expected loss we assume that the prior distribution of $(b_i, 1/\sigma_i^2)$ is normal-gamma (see the appendix) where $h_i \equiv 1/\sigma_i^2$ and the parameter of the prior is $(m', \mathrm{v}', n', \nu')$. Thus one might use previous estimates of b_i to form $(m', \mathrm{v}', n', \nu')$ or use $m' = \bar{b}$, $\mathrm{v}' = \bar{\sigma}^2$. As mentioned in the appendix, the posterior distribution will also be normal-gamma with parameters $(m'', \mathrm{v}'', n'', \nu'')$; the unprimed parameters reflect the sampling-distribution values. Since we are most interested in the posterior distributions associated with b, we will use the marginal distribution of $f_{N\gamma}(b,h|m'', \mathrm{v}'', n'', \nu'')$ which (see appendix) is the Student's distribution $g(b|m'', \mathrm{v}'', n'', \nu'')$. Therefore the expected loss is

$$E_L(n) = \int \int \int L(b, m'')\, g(b|m'', \mathrm{v}'', n'', \nu'')\, D(m, \mathrm{v}|m', \mathrm{v}', n', \nu'; n, \nu)\, dm\, d\mathrm{v}\, db,$$

where $D(\cdot)$ is the sampling distribution (see appendix) and the updating equalities above are used. Finally, this means that the optimal sample size n^* is the number of data points that solves:

$$\min_n E_L(n) + c(n)$$

Evaluating $E_L(n)$, while tedious, will not be difficult, especially since $L(b, m'')$ is piecewise-linear.

OTHER ISSUES

First, it should be obvious that a similar aproach can be taken, via duality, for changes in one element of the price vector p. More important, however, is the problem of restricting the analysis to one element of the parameter set, be it b or p. One would normally expect to see a sensitivity discussion which considered changes in the entire vector, not just one element.

A moment's reflection will bring to light why this has not been pursued. In standard sensitivity analysis the right-hand side is parameterized on a *scalar* and ranging is performed by varying the scalar. This amounts to requiring the right-hand side to expand along a ray, which does not make sense in the present application. On the other hand, while the loss function for such a procedure is well understood and easy to compute for various values (this being a direct extension of the material in the previous sections), the loss function to be used for the problem at hand (where b would not expand along a ray) is not known. In fact, we would not have the benefit of not having to solve the program for each $\hat{b}$, but would essentially have to solve an infinite number of programs in order to find the optimal $\hat{b}$.

The obvious heuristic is to vary one element at a time. This is clearly not optimal. It should be clear that each element of the right-hand side would be a function of all other elements and that, for example, the "optimal" $\hat{b}$ would be sequence dependent, i.e., would change depending on the sequence of $\hat{b}_i$ examined.

SUMMARY

This paper is an attempt to apply basic decision theory to sensitivity questions in linear programming. We have taken the position that a solution is sensitive only if the expected gain

from improving the solution exceeds the cost of improvement. This would seem to be an obvious and intuitive approach, but it does not seem to be the approach commonly used in sensitivity analysis.†

While closed form solutions do not appear possible, locally optimal solutions are clearly possible if the user has the facility to do basic sensitivity analysis and has access to (or will acquire) a computer routine for numerical integration.

ACKNOWLEDGMENTS

The author expresses thanks to V. Balachandran, D. P. Baron, S. Deshmukh, S. Meyer and D. J. Roberts, who are in no way responsible for remaining errors.

REFERENCES

[1] Cunningham, A. A., and D. M. Frances, "A Data Collection Strategy for Estimation of Cost Coefficients of a Linear Programming Model" *Management Science*, **22**, No. 10, 1074-1080 (1976).

[2] Raiffa, H., and R. Schlaiffer, *Applied Statistical Decision Theory*, (MIT Press, Cambridge, 1961).

APPENDIX

The following is shown in [2]:

1. If h is distributed Gamma-2 then we mean that

$$f_{\gamma^2}(h|\mathrm{v},\ \nu) = e^{-\mathrm{v}\nu h/2}\left(\frac{1}{2}\,\mathrm{v}\nu h\right)^{\frac{\nu}{2}-1}\frac{\mathrm{v}\nu/2}{\Gamma\left(\frac{\nu}{2}\right)}.$$

Thus, if $(\mu,\ h)$ is distributed Normal-gamma then $f_{N\gamma}(\mu,\ h|m,\ \mathrm{v},\ n,\ \nu) = f_N(\mu|m,\ hn) f_{\gamma^2}(h|\mathrm{v},\ \nu)$.

2. If the prior distribution of $(\mu,\ h)$ is Normal-gamma with parameter $(m',\ \mathrm{v}',\ n',\ \nu')$ and if a sample then yields a sufficient statistic $(m,\ \mathrm{v},\ n,\ \nu)$, the posterior distribution of $(\mu,\ h)$ is Normal-gamma with parameters $(m'',\ \mathrm{v}'',\ n'',\ \nu'')$:

$$m'' = (n'm' + nm)/(n' + n),$$

$$n'' = n' + n,$$

$$\nu'' = \nu' + \nu + \delta(n') + \delta(n) - \delta(n''),$$

$$\mathrm{v}'' = ([\nu'\mathrm{v}' + n'(m')^2] + [\nu\mathrm{v} + nm^2] - n''(m'')^2)/\nu'',$$

where

$$\delta(x) = \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

†A recent exception is [1].

3. If the joint distribution of (μ, h) is Normal-gamma as defined above, then the marginal distribution of μ is the Student distribution

$$g(\mu|m, \mathrm{v}, n, \nu) = \frac{\nu^{\nu/2}\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1\nu}{2}\right)} \left[\nu + \frac{n}{\mathrm{v}}(\mu - m)^2\right]^{-n/2} \sqrt{\frac{n}{\mathrm{v}}}.$$

4. If a sample of size $n > 0$ is to be taken from an independent Normal process with parameter (μ, h) having a Normal-gamma distribution with parameter $(m', \mathrm{v}', n', \nu')$, with $(\mathrm{v}', n', \nu' > 0)$, then the unconditional joint distribution of the statistic (m, v) is

$$D(m, \mathrm{v}|m', \mathrm{v}', n'; n, \nu) = A(n) \frac{(\nu\mathrm{v})^{\nu/2-1}}{(\nu'\mathrm{v}' + \nu\mathrm{v} + n_u [m - m']^2)^{\nu''/2}},$$

where

$$n_u = n'n/(n + n'),$$

$$\nu'' = \nu' + \nu + 1,$$

$$A(n) = \frac{\Gamma(\nu''/2)}{\Gamma\left(\frac{n-1}{2}\right)} \sqrt{\frac{n}{n+n'}\left(\frac{n-1}{2}\right)^{n-1}} \cdot \sqrt{\frac{n'}{2\pi}} \cdot \frac{\left(\frac{\mathrm{v}'\nu'}{2}\right)^{\nu'/2}}{\Gamma(\nu'/2)},$$

$\Gamma(x)$ = gamma function of x.

TWO QUEUEING SYSTEMS SHARING THE SAME FINITE WAITING ROOM

M. J. Fischer

Defense Communications Engineering Center
Reston, Virginia

ABSTRACT

In this paper, we consider the analysis of two M/G/1 queueing systems sharing the same finite waiting room. An exact analysis is given for several special cases, and then an algorithm is developed which approximates the system behavior for the general problem. Comparisons are made between the special cases and the algorithm.

1. INTRODUCTION

In this paper we consider the case in which two classes of traffic arrive at the same finite waiting room (K spaces). There are two servers, one for each type of customer. A server is allowed to service one and only one class of customer. We assume that the customers arrive in accordance with two independent Poisson processes with parameters λ_i, $i = 1,2$. The random variables S_i, representing the time to service a class i customer, are assumed to have a general distribution function, which is independent of each other and of the arrival processes. If the server for a given class of customer is busy upon the customer's arrival and all the K spaces are full, the customer leaves the system without receiving service. We will be considering the system in the steady state.

The case of only one class of customers has been considered previously [3,4,6] and recently there has been some additional interest in the single-server case [5,7]. However, to the best of our knowledge no work has been done on the case where there are two classes of customers and a server for each class.

Such a system could serve as a model for the buffer in a data telecommunications system. In designing a data telecommunications network, the buffers (waiting rooms) are usually sized for a negligible probability of overflow [8]. The effect of buffer size (where the entire buffer contents are served by a single server) has been considered [1,2]. However, in a real system a buffer may store messages (customers) for several nodes (servers). Thus, the buffer is a shared resource for messages destined for these nodes and the service each type of message receives depends on the level of traffic of the other messages using the buffer.

For the system considered in this paper, the analysis of the general-service-distribution case is complicated and not amenable to a straightforward analysis. In section 2, we give an analysis for some special cases. An algorithm which approximates the steady-state behavior of the system with general service-time distributions is presented in section 3. Comparisons with the results obtained in section 2 are given at the end of section 3 and in the numerical examples considered in section 4.

2. SOME SPECIAL CASES

Three special cases are considered in this section. Let Q_i, $i = 1, 2$, be the steady-state number of class i customers in the system (in service and waiting), and define $P_{i,j} = \Pr\{Q_1 = i, Q_2 = j\}$. We note when $i = 0, 1$, then $P_{i,j} \equiv 0$ for $j \geq K + 2$, and when $2 \leq i \leq K + 1$, then $P_{i,j} \equiv 0$ for $j \geq K + 3 - i$. Within a class, service is rendered on a first-come, first-served basis.

a. $K = 0$. For the case where there are no waiting spaces (K = 0) each class of customer behaves independently of the other and as a single-server loss system. Thus, if $\rho_i = \lambda_i E\{S_i\}$ $(i = 1, 2)$, then, for $i = 0, 1$ and $j = 0, 1$, we have

$$P_{i,j} = \frac{\rho_1^{\,i}\,\rho_2^{\,j}}{(1 + \rho_1)\,(1 + \rho_2)}. \tag{1}$$

b. *Exponential Service for Both Classes.* For the case where each class has an exponential service distribution, with $E\{S_i\} = 1/\mu_i$, $i = 1, 2$, the steady-state equations for $P_{i,j}$ are

$$(\lambda_1 + \lambda_2)\,P_{0,0} = \mu_1 P_{1,0} + \mu_2 P_{0,1}, \tag{2}$$

$$\vdots$$

$$(\lambda_1 + \lambda_2 + \mu_2)\,P_{0,j} = \mu_1 P_{1j} + \lambda_2 P_{0,j-1} + \mu_2 P_{0,j+1},\quad 1 \leq j \leq K,$$

$$\vdots$$

$$(\lambda_1 + \mu_2)\,P_{0,K+1} = \mu_1 P_{1,K+1} + \lambda_2 P_{0,K},$$

$$(\lambda_1 + \lambda_2 + \mu_1)\,P_{i,0} = \lambda_1 P_{i-1,0} + \mu_1 P_{i+1,0} + \mu_2 P_{i,1},$$

$$\vdots$$

$$(\lambda_1 + \lambda_2 + \mu_1 + \mu_2)\,P_{i,j} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + \mu_1 P_{i+1,j} + \mu_2 P_{i,j+1},\quad \begin{cases} i = 1, 2, \ldots, K \\ j = 1, 2, \ldots, K + 1 - i, \end{cases}$$

$$\vdots$$

$$(\mu_1 + \mu_2)\,P_{i,K+2-i} = \lambda_1 P_{i-1,K+2-i} + \lambda_2 P_{i,K+1-i},$$

$$(\lambda_2 + \mu_1)\,P_{K+1,0} = \lambda_1 P_{K,0} + \mu_2 P_{K+1,1},$$

$$(\mu_1 + \mu_2)\,P_{K+1,1} = \lambda_1 P_{K,1} + \lambda_2 P_{K+1,1}.$$

The solution to these equations is ($\rho_i = \lambda_i/\mu_i$)

$$P_{i,j} = P_{00}\rho_1^i\rho_2^j, \tag{3}$$

where

$$P_{0,0} = \frac{(1-\rho_1)(1-\rho_2)(\rho_2-\rho_1)}{(1-\rho_1^{K+2})(\rho_2-\rho_1)-\rho_2(1-\rho_1)(\rho_2^{K+2}-\rho_1^{K+2})+\rho_1\rho_2(1-\rho_1)(1-\rho_2)(\rho_2^{K+1}-\rho_1^{K+1})}$$

From equation (3) the marginal probability distributions of Q_1 and Q_2 are obtained in a standard manner. We present results for Q_1 (the analogous results can be obtained for Q_2 via symmetry):

$$\Pr\{Q_1 = i\} = \begin{cases} P_{0,0}\dfrac{1-\rho_2^{K+2}}{1-\rho_2}, & i = 0, \\ P_{0,0}\rho_1^i\dfrac{1-\rho_2^{K+3-i}}{1-\rho_2}, & i = 1, 2, \ldots, K+1. \end{cases} \tag{4}$$

From (4) we get the expected number of class-1 customers in the system $E\{Q_1\}$ to be

$$E\{Q_1\} = \frac{P_{0,0}}{1-\rho_2}\left[\rho_1\left[\frac{(1-\rho_1^{K+2})-(K+2)(1-\rho_1)\rho_1^{K+1}}{(1-\rho_1)^2}\right]\right. \tag{5}$$
$$\left.-\rho_1\rho_2^2\left[\frac{(\rho_2^{K+2}-\rho_1^{K+2})-(K+2)\rho_1^{K+1}(\rho_2-\rho_1)}{(\rho_2-\rho_1)^2}\right]\right].$$

Another measure of performance for class 1 customers is the probability of blocking upon arrival, PB_1; this quantity is given by

$$PB_1 = P_{K+1,0} + \sum_{i=1}^{K+1} P_{i,K+2-i} \tag{6}$$

$$= P_{00}\left[\frac{\rho_1^{K+1}(\rho_2-\rho_1)+\rho_1\rho_2(\rho_2^{K+1}-\rho_1^{K+1})}{(\rho_2-\rho_1)}\right].$$

It is interesting to note that the probability of blocking for class 2 customers is not the same as PB_1, even though their arrival processes are Poisson. This stems from the fact that the states of the system under which customers are blocked are different for different classes of traffic. A final measure of performance for the class 1 customer, the expected waiting time $E\{W_1\}$, is given by Little's Formula

$$E\{W_1\} = E\{Q_1\}/(1-PB_1)\,\lambda_1. \tag{7}$$

c. *$K = 1$, Class 1 Service Exponential, Class 2 Service Erlang 2.* As mentioned in section 1, the analysis of the case of general service times is not straightforward. Using standard methods of analyzing such systems, i.e., supplementary variables or embedded Markov Chain, does not yield results directly. It does appear, however, that the case of Erlangian or hyperexponential service-time distributions can be handled using the method of phases [9]. We will

demonstrate how one could go about the solution of this system for the special case of K = 1, exponential service distribution for class 1, $E\{S_1\} = 1/\mu_1$, and an Erlangian service distribution with shape parameter equal to 2 ($E\{S_2\} = 2/\mu_2$) for class 2. Thus, every class 2 customer who is accepted into the system brings in two phases.

Let R_{ij} represent the probability of $Q_1 = i$, and let the number of class 2 phases in the system equal j. The steady-state equations for the nonzero values of R_{ij} are

$$
\begin{aligned}
(\lambda_1 + \lambda_2)\, R_{00} &= \mu_1 R_{10} + \mu_2 R_{01}\,, \\
(\lambda_1 + \lambda_2 + \mu_2)\, R_{01} &= \mu_1 R_{10} + \mu_2 R_{02}\,, \\
(\lambda_1 + \lambda_2 + \mu_2)\, R_{02} &= \mu_1 R_{12} + \lambda_2 R_{00} + \mu_2 R_{03}\,, \\
(\lambda_1 + \mu_2)\, R_{03} &= \mu_1 R_{13} + \lambda_2 R_{01} + \mu_2 R_{04}\,, \\
(\lambda_1 + \mu_2)\, R_{04} &= \mu_1 R_{14} + \lambda_2 R_{02}\,;
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
(\lambda_1 + \lambda_2 + \mu_1)\, R_{10} &= \mu_1 R_{20} + \mu_2 R_{11} + \lambda_1 R_{00}\,, \\
(\lambda_1 + \lambda_2 + \mu_1 + \mu_2)\, R_{11} &= \lambda_1 R_{01} + \mu_1 R_{21} + \mu_2 R_{12}\,, \\
(\lambda_1 + \lambda_2 + \mu_1 + \mu_2)\, R_{12} &= \lambda_1 R_{02} + \mu_1 R_{22} + \lambda_2 R_{10} + \mu_2 R_{13}\,, \\
(\mu_1 + \mu_2)\, R_{13} &= \lambda_1 R_{03} + \lambda_2 R_{11} + \mu_2 R_{14}\,, \\
(\mu_1 + \mu_2)\, R_{14} &= \lambda_1 R_{04} + \lambda_2 R_{12}\,;
\end{aligned} \tag{9}
$$

and

$$
\begin{aligned}
(\lambda_2 + \mu_1)\, R_{20} &= \lambda_1 R_{10} + \mu_2 R_{21}\,, \\
(\mu_1 + \mu_2)\, R_{21} &= \lambda_1 R_{11} + \mu_2 R_{22}\,, \\
(\mu_1 + \mu_2)\, R_{22} &= \lambda_1 R_{12} + \lambda_2 R_{22}\,.
\end{aligned} \tag{10}
$$

Now let $\eta_i = \lambda_i/\mu_i$ ($i = 1, 2$), $\alpha = \mu_2/\mu_1$, and

$$
\bar{R}_0 = \begin{bmatrix} R_{00} \\ R_{01} \\ R_{02} \\ R_{03} \\ R_{04} \end{bmatrix}, \quad \bar{R}_1 = \begin{bmatrix} R_{10} \\ R_{11} \\ R_{12} \\ R_{13} \\ R_{14} \end{bmatrix}, \quad \text{and } \bar{R}_2 = \begin{bmatrix} R_{20} \\ R_{21} \\ R_{22} \\ 0 \\ 0 \end{bmatrix}.
$$

From equation (8) we get the following matrix equation:

$$
\bar{R}_1 = A_1 \bar{R}_0 \tag{11}
$$

where

$$
A_1 = \begin{bmatrix}
\eta_1+\alpha\eta_2 & -\alpha & 0 & 0 & 0 \\
0 & \eta_1+\alpha\eta_2+\alpha & -\alpha & 0 & 0 \\
-\alpha\eta_2 & 0 & \eta_1+\alpha\eta_2+\alpha & -\alpha & 0 \\
0 & -\alpha\eta_2 & 0 & \eta_1+\alpha & -\alpha \\
0 & 0 & -\alpha\eta_2 & 0 & \eta_1+\alpha
\end{bmatrix};
$$

and from the first three equations in (9) and equation (11)

$$\bar{R}_2 = A_2 \bar{R}_1 - B_2 \bar{R}_0 \tag{12}$$
$$= (A_2 A_1 - B_2) \bar{R}_0 ,$$

where

$$A_2 = \begin{bmatrix} \eta_1+\alpha\eta_2+1 & -\alpha & 0 & 0 & 0 \\ 0 & \eta_1+\alpha\eta_2+1+\alpha & -\alpha & 0 & 0 \\ -\alpha\eta_2 & 0 & \eta_1+\alpha\eta_2+1+\alpha & -\alpha & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$B_2 = \begin{bmatrix} \eta_1 & 0 & 0 & 0 & 0 \\ 0 & \eta_1 & 0 & 0 & 0 \\ 0 & 0 & \eta_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

From equations (11) and (12) one can see that the solution rests on determining the five unknowns, R_{0j}, $j = 0, 1, 2, 3, 4$. Thus we need to generate five independent equations in these unknowns. From (8), (9), and (10) there are still five equations (the last two from (9) and the three from equation (10)), but one of them is redundant. Dropping the fourth equation of (9) gives us four independent equations; the final one can be found from the normalizing condition $\sum_{i,j} R_{ij} = 1$. Denote $C_1 = A_1$, and $C_2 = A_2 A_1 - B_2$ and $C_l(i,j)$ as the i,j th entry of C_l; then from equations (11) and (12) for $j = 0, 1, \ldots, 4$ we have

$$R_{lj} = \sum_{n=0}^{4} C_l(j,n) R_{0n}, \quad l = 1,2. \tag{13}$$

The required five equations now become

$$\sum_{n=0}^{4} [(1+\alpha) C_1(4,n) - \alpha \eta_2 C_1(2,n)] R_{0n} - \eta_1 R_{04} = 0 ,$$

$$\sum_{n=0}^{4} [(\alpha\eta_2 + 1) C_2(0,n) - \alpha C_2(1,n) - \eta_1 C_1(0,n)] R_{0n} = 0 ,$$

$$\sum_{n=0}^{4} [(1+\alpha) C_2(1,l) - \alpha C_2(2,l) - \eta_1 C_1(1,l)] R_{0n} = 0 , \tag{14}$$

$$\sum_{n=0}^{4} [(1+\alpha) C_2(2,n) - \eta_2 \alpha C_2(0,n) - \eta_1 C_1(2,l)] R_{0n} = 0 ,$$

$$\sum_{n=0}^{4} \left[1 + \sum_{i=1}^{2} \sum_{j=0}^{4} C_i(j,n)\right] R_{0n} = 1 .$$

Thus, the solution to this special case rests on solving the five independent equations given by (14). It appears that one could generalize the method just described for solution of the case of Erlangian service distribution and general K. But use of this method would require solution of a large system of equations. In the next section we present an algorithm which can be used to approximate quickly and efficiently the steady state performance for general service-time distributions.

3. AN ALGORITHM APPROXIMATING THE GENERAL SERVICE TIME DISTRIBUTION CASE

The algorithm approximating the case of general service-time distributions is based on the following observation: When there are $j(\geqslant 1)$ class 2 customers in the system until the next class 2 arrival or departure, whichever is first, the system behaves like an M/G/1/K+1-j system for the class 1 customers. Analogously, the same sort of phenomenon happens with respect to the class 2 customers.

Now, from the definition of conditional probability, we have

$$\Pr\{Q_1 = i | Q_2 = j\} \Pr\{Q_2 = j\} = \Pr\{Q_1 = i,\ Q_2 = j\} \tag{15}$$

$$= \Pr\{Q_2 = j | Q_1 = i\} \Pr\{Q_1 = i\},$$

or

$$\Pr\{Q_1 = i\} = \frac{\Pr\{Q_1 = i | Q_2 = j\}}{\Pr\{Q_2 = j | Q_1 = i\}} \Pr\{Q_2 = j\}; \tag{16}$$

but $\sum_{i=0}^{K+1} \Pr\{Q_1 = i\} = 1$, and so for $j = 0, 1, 2, \ldots, K+1$

$$\Pr\{Q_2 = j\} = \left[\sum_{i=0}^{K+1} \frac{\Pr\{Q_1 = i | Q_2 = j\}}{\Pr\{Q_2 = j | Q_1 = i\}}\right]^{-1}. \tag{17}$$

Analogously, for $i = 0, 1, 2, \ldots, K+1$

$$\Pr\{Q_1 = i\} = \left[\sum_{j=0}^{K+1} \frac{\Pr\{Q_2 = j | Q_1 = i\}}{\Pr\{Q_1 = i | Q_2 = j\}}\right]^{-1}. \tag{18}$$

The algorithm uses equations (17) and (18) to generate an approximation to $\Pr\{Q_1 = i,\ Q_2 = j\}$. However, the conditional probabilities $\Pr\{Q_1 = i | Q_2 = j\}$ and $\Pr\{Q_2 = j | Q_1 = i\}$ are not known. As stated earlier, when $Q_2 = j$ the system instantaneously behaves as an M/G/1/K+1-j system for the class 1 customers, and so we use known results for these systems to approximate $\Pr\{Q_1 = i | Q_2 = j\}$ (see [3] or [4]). That is, when $Q_2 = j$ we use the state probabilities obtained from an M/G/1/K+1-j system where only class 1 customers are arriving. We note when $j = 0$ and 1 the conditional probabilities are obtained from the state probabilities for an M/G/1/K queueing system. Analogously, $\Pr\{Q_2 = j | Q_1 = i\}$ are approximated using the class 2 results from an M/G/1/K+1-i system. Let us denote these approximations by $\hat{P}_{i|Q_2=j}$ and $\hat{P}_{j|Q_1=i}$, and all other approximate probabilities by the use of a circumflex ($\hat{}$).

The algorithm first approximates $\Pr\{Q_1 = i,\ Q_2 = j\}$ by conditioning on the number of class 2 customers in the system. Since class 2 results are being used, we denote this approximation by $\hat{P}_{i,j}^{(2)}$; it is given by

$$\hat{P}_{i,j}^{(2)} = \hat{P}_{i|Q_2=j} \hat{\Pr}\{Q_2 = j\}, \tag{19}$$

where

$$\hat{\Pr}\{Q_2 = j\} = \left[\sum_{i=0}^{K+1} \frac{\hat{P}_{i|Q_2=j}}{\hat{P}_{j|Q_1=i}}\right]^{-1}. \tag{20}$$

For $j = 0$ and $j = 1$ we use equation (20) to determine $\hat{\Pr}\{Q_2 = j\}$. Once these quantities have been obtained, $\hat{P}_{i,j}^{(2)}$ (for $i = 0, 1, 2, \ldots, K+1$ and $j = 0, 1$) can be found from equation (19). When $Q_2 = 2$, $\hat{P}_{K+1|Q_2=2}$ and $\hat{P}_{2|Q_1=K+1}$ do not make sense, but from the definition of conditional probability we have

$$\frac{\hat{P}_{K+1|Q_2=2}}{\hat{P}_{2|Q_1=K+1}} = \frac{\hat{\Pr}\{Q_1 = K+1\}}{\hat{\Pr}\{Q_2 = 2\}}. \tag{21}$$

Using equation (20) we now get

$$\hat{\Pr}\{Q_2 = 2\} = \frac{1 - \hat{\Pr}\{Q_1 = K+1\}}{\sum_{i=0}^{K} \frac{\hat{P}_{i|Q_2=2}}{\hat{P}_{2|Q_1=i}}}. \tag{22}$$

Since for $j = 0$ and $j = 1$ we have generated $\hat{P}_{K+1,0}^{(2)}$ and $\hat{P}_{K+1,1}^{(2)}$; we have

$$\hat{\Pr}\{Q_1 = K+1\} = \hat{P}_{K+1,0}^{(2)} + \hat{P}_{K+1,1}^{(2)},$$

and so

$$\hat{\Pr}\{Q_2 = 2\} = \frac{1 - \hat{P}_{K+1,0}^{(2)} - \hat{P}_{K+1,1}^{(2)}}{\sum_{i=0}^{K} \frac{\hat{P}_{i|Q_2=2}}{\hat{P}_{2|Q_1=i}}}. \tag{23}$$

Using equations (19) and (23), we now can approximate $\hat{P}_{i,2}$ for $i = 0, 1, 2, \ldots, K$. For $j \geqslant 3$, we can sequentially develop an expression for $\hat{\Pr}\{Q_2 = j\}$ using equation (20) and the results previously generated for $\hat{P}_{i,n}^{(2)}$, $n = 0, 1, 2, \ldots, j-1$. For $j = 2, \ldots, K$, one gets

$$\hat{\Pr}\{Q_2 = j\} = \frac{1 - \sum_{n=K+3-j}^{K+1} \sum_{l=0}^{K+2-n} \hat{P}_{n,l}^{(2)}}{\sum_{i=0}^{K+2-j} \frac{\hat{P}_{i|Q_2=j}}{\hat{P}_{j|Q_1=i}}}; \tag{24}$$

and $\hat{\Pr}\{Q_2 = K+1\} = 1 - \sum_{j=0}^{K} \hat{\Pr}\{Q_2 = j\}$.

In a similar fashion we can use equations (18) and (19) to generate an approximation to $\Pr\{Q_1 = i, Q_2 = j\}$ conditioned on the class 1 customers in the system, denoted by $\hat{P}_{ij}^{(1)}$. In general, it turns out that $\hat{P}_{ij}^{(1)} \neq \hat{P}_{ij}^{(2)}$, and the final approximation to $\Pr\{Q_1 = i, Q_2 = j\}$, denoted by $\hat{P}_{i,j}$, is given by

$$\hat{P}_{i,j} = \frac{\rho_1 \hat{P}_{i,j}^{(1)} + \rho_2 \hat{P}_{i,j}^{(2)}}{\rho_1 + \rho_2}, \tag{25}$$

where $\rho_i = \lambda_i E(S_i)$, $i = 1, 2$.

The steps of the algorithm can now be summarized:

STEP 1: Use equation (20) to determine $\hat{\Pr}\{Q_2 = j\}$ for $j = 0, 1$. Use equation (19) to determine $\hat{P}_{i,j}^{(2)}$ for $j = 0, 1$ and $i = 0, 1, \ldots, K + 1$.

STEP 2: For $j = 2, 3, \ldots, K$ use equation (24) to determine $\hat{\Pr}\{Q_2 = j\}$ and equation (19) to determine $\hat{P}_{ij}^{(2)}$ for $i = 0, 1, \ldots K + 2 - j$.

STEP 3: Set $\hat{\Pr}\{Q_2 = K + 1\} = 1 - \sum_{j=0}^{K} \hat{\Pr}\{Q_2 = j\}$ and use equation (19) to determine $\hat{P}_{0,K+1}^{(2)}$ and $\hat{P}_{1,K+1}^{(2)}$.

STEP 4: Return to step 1 to determine the approximation to $\Pr\{Q_1 = i, Q_2 = j\}$ based on class 1, $\hat{P}_{i,j}^{(1)}$.

STEP 5: Once $\hat{P}_{i,j}^{(1)}$ and $\hat{P}_{i,j}^{(2)}$ have been determined set $\hat{P}_{i,j} = \dfrac{\rho_1 \hat{P}_{i,j}^{(1)} + \rho_2 \hat{P}_{i,j}^{(2)}}{\rho_1 + \rho_2}$.

Once $\hat{P}_{ij}$ has been determined, all the desired measures of performance can be obtained by the use of standard results.

Several comments regarding the accuracy of the algorithm can be made. For the case where $K = 0$, the algorithm is exact. We have, for $i = 0, 1$ and $j = 0, 1$,

$$\hat{P}_{i|Q_2=j} = \frac{\rho_1^{\,i}}{1 + \rho_1} \tag{26}$$

and

$$\hat{P}_{j|Q_1=i} = \frac{\rho_2^{\,j}}{1 + \rho_2} \tag{27}$$

Using equation (20), for $j = 0, 1$, we get

$$\begin{aligned}\hat{\Pr}\{Q_2 = j\} &= \left[\sum_{i=0}^{1} \frac{\hat{P}_{i|Q_2=1}}{\hat{P}_{j|Q_1=i}}\right]^{-1} \\ &= \left[\sum_{i=0}^{1} \frac{\rho_1^{\,i}}{1 + \rho_1} \frac{1 + \rho_2}{\rho_2^{\,j}}\right]^{-1} \\ &= \frac{\rho_2^{\,j}}{1 + \rho_2};\end{aligned}$$

so, for $i = 0, 1$ and $j = 0, 1$,

$$(29)\quad \hat{P}_{i,j}^{(2)} = \frac{\rho_1^{\,i}\rho_2^{\,j}}{(1+\rho_1)(1+\rho_2)} = \frac{\rho_1^{\,i}\rho_2^{\,j}}{1+\rho_1+\rho_2+\rho_1\rho_2},$$

which is equation (1). Analogously, one can see that $\hat{P}_{i,j}^{(1)} = \hat{P}_{i,j}^{(2)}$, and so the algorithm generates the exact results for the case when $K = 0$.

For the case of exponential service-time distributions and general K, section 2b, it can be shown that $\hat{P}_{i,j}^{(2)} \equiv \hat{P}_{i,j}^{(1)}$ and the algorithm generates exact results. Using the results of section 2b to obtain an expression for $\Pr\{Q_2 = j | Q = i\}$ (for $i \geqslant 1$ and $j \geqslant 1$), we have

$$(30)\quad \Pr\{Q_2 = j | Q_1 = i\} = \frac{\Pr\{Q_1 = i,\ Q_2 = j\}}{\Pr\{Q_1 = i\}} = \frac{P_{00}\rho_1\rho_2}{P_{00}\rho_1^{\,i}\left(\dfrac{1-\rho_2^{\,k+3-i}}{1-\rho_2}\right)} = \rho_2^{\,j}\frac{(1-\rho_2)}{1-\rho_2^{\,K+3-i}}.$$

But equation (30) is exactly the result we would have obtained for $\hat{P}_{j|Q_1=i}$ by using the state probabilities for an M/M/1/K+1-i. Thus, $\hat{P}_{j|Q_1=i}$ is exact and can be used along with quite a bit of algebra to show that the algorithm generates exact results for this case.

In the next section we give numerical comparisons of the algorithm for the situation where $K > 0$ and service distribution is not exponential. There the algorithm is not exact, but it performs very well.

4. NUMERICAL RESULTS

In the first part of this section we give some numerical comparisons of the algorithm for some situations where $K \geqslant 1$ and the service distributions are not both exponentially distributed. The section closes by using the results of the analysis of section 2b to answer the question as to whether or not to have separate waiting facilities for each class.

Table 1 is a comparison of the algorithm presented in section 3 and the special case of K = 1, one service-time distribution exponential, and the other E_2 (section 2c). For a fixed combined load of 0.9, comparisons between the exact (E) and the approximate (A) results are presented for various measures of performance. It is immediately discernible that the algorithm is quite accurate for all the cases presented. In a relative sense, the algorithm seems to be worse when the traffic intensities are equal. Furthermore, when one of the traffic intensities is zero the algorithm gives the exact results.

TABLE 1. *Comparison of algorithm and the case where one service time distribution is exponential and the other Erlangian* ($K = 1$, $\mu_2 = 2\mu_1$)

Traffic Intensities			$E(Q_1)$	$E(Q_2)$	PB_1	PB_2	COV (Q_1, Q_2)
1	2	*					
0	0.9	E	0.	0.92731	0.	0.27306	0.
		A	0.	0.92731	0.	0.27306	0.
0.1	0.8	E	0.10407	0.84445	0.02824	0.23825	-0.00511
		A	0.10406	0.84470	0.02827	0.23851	-0.00474
0.2	0.7	E	0.21437	0.75031	0.05825	0.20491	-0.01563
		A	0.21433	0.75100	0.5835	0.20557	-0.01474
0.3	0.4	E	0.32799	0.64783	0.08987	0.17306	-0.02557
		A	0.32789	0.64884	0.09001	0.17391	-0.02456
0.4	0.5	E	0.44204	0.53987	0.12288	0.14264	-0.03098
		A	0.44189	0.54095	0.12303	0.14334	-.03033
0.45	0.45	E	0.49834	0.48469	0.13982	0.12789	-0.03170
		A	0.49819	0.48571	0.13997	0.12842	-0.03101
0.5	0.4	E	0.55373	0.42914	0.15702	0.11343	-0.03017
		A	0.55358	0.43003	0.15714	0.11374	-0.03011
0.6	0.3	E	0.66049	0.31804	0.19199	0.08510	-0.02363
		A	0.66038	0.31861	0.19207	0.08502	-0.02405
0.7	0.2	E	0.76002	0.20857	0.22750	0.05720	-0.01373
		A	0.75997	0.20881	0.22753	0.5693	-0.01426
0.8	0.1	E	0.85036	0.10223	0.26323	0.02909	-0.00429
		A	0.85035	0.10227	0.26323	0.02909	-0.00454
0.9	0	E	0.92989	0.	0.29889	0.	0.
		A	0.92989	0.	0.29889	0.	0.

*E = exact; A = approximate

In Table 2 a comparison of the algorithm and the results of a simulation are given for the case K = 5, with one service distribution E_2 and the other E_4. For various combinations of loads, comparisons between the simulation (S) and the approximate (A) results are presented for various measures of performance. Although the approximation does very well, it does not appear to be as accurate as in the previous comparisons. To some extent, this may be attributed to the fact that this comparison was against a simulation, with its own inherent inaccuracies. However, the results are certainly representative and the approximation appears to be quite suitable for most applied work.

TABLE 2. *Comparison of algorithm and the case where one service-time distribution is* E_2 *and the other* E_4 ($K = 5$, $E(S_1) = E(S_2) = 1$)

Traffic Intensities		*	$E(Q_1)$	$E(Q_2)$	PB_1	PB_2
1	2					
0.1	0.1	S	0.117	0.103	0.000	0.000
		A	0.108	0.108	0.000	0.000
0.5	0.1	S	0.857	0.108	0.003	0.000
		A	0.854	0.107	0.003	0.000
0.9	0.1	S	2.557	0.114	0.081	0.005
		A	2.539	0.112	0.077	0.006
0.1	0.9	S	0.103	2.656	0.005	0.073
		A	0.107	2.545	0.007	0.068
0.5	0.9	S	0.807	2.350	0.038	0.070
		A	0.766	2.416	0.040	0.074
0.9	0.9	S	2.130	1.980	0.128	0.120
		A	2.055	2.032	0.121	0.119

*S = simulation; A = approximate

Part of the reason for investigating this queueing system was to try to answer the following question: "Are there conditions under which it is better to give each class of customer his own waiting spaces and not allow the other class to use them?" We have used the results obtained for the exponential service-distribution case, section 2b, to investigate this question numerically. Table 3 gives the results of this investigation. For different values of K, three basic cases were considered: $\rho_i \geqslant 1$, $i = 1, 2$; $\rho_1 < 1$, $\rho_2 \geqslant 1$; and $\rho_i < 1$, $i = 1, 2$. For ρ_1, ρ_2 and K fixed, the total load carried for the case where both classes are allowed to use all K spaces (same) is compared with the load carried for the case where the K spaces are optimally divided among the classes and a customer may only use the spaces allowed to his class (sep.). The optimal allocation of the K spaces to each class (given in parentheses) was found by formulating the problem as a dynamic programming program, where the objective function was to maximize the total load carried for the system. From the table, one sees that if

both traffic intensities are greater than 1, use of separate waiting facilities for each class is optimal, and for the case where one class's traffic intensity is less than 1 and the other's greater than 1, the solution depends on K and on the traffic intensities. For the case where both traffic intensities are less than 1, it is better for both classes to use the same spaces. It should be pointed out that a more detailed study of the optimal operation of this system has been given by Latouche [10].

TABLE 3. *Comparison of same or separate waiting facilities for each class of customers*

		Total load carried for separate and same spaces (and optimal allocation for separate spaces)							
		K = 3		K = 5		K = 7		K = 9	
ρ_1	ρ_2	Sep.	Same	Sep.	Same	Sep.	Same	Sep.	Same
1.1	1.9	1.6310 (2,1)	1.6150	1.7610 (3,2)	1.6820	1.8330 (4,3)	1.7070	1.8750 (5,4)	1.7170
1.2	1.8	1.6480 (2,1)	1.6390	1.7810 (3,2)	1.7170	1.855 (4,3)	1.7520	1.8980 (5,4)	1.7690
1.3	1.7	1.6590 (2,1)	1.6560	1.7940 (3,2)	1.7440	1.8690 (4,3)	1.7890	1.9130 (5,4)	1.8140
1.4	1.6	1.6650 (2,1)	1.6670	1.8010 (3,2)	1.7620	1.8750 (4,3)	1.8150	1.9200 (5,4)	1.8470
1.5	1.5	1.6650 (2,1)	1.6690	1.8010 (3,2)	1.7680	1.8760 (4,3)	1.8240	1.9210 (5,4)	1.8590
0.1	1.7	1.0380 (0,3)	1.0420	1.0730 (0,5)	1.0780	1.0890 (1,6)	1.0900	1.0960 (1,8)	1.0940
0.3	1.5	1.1570 (1,2)	1.1890	1.2320 (1,4)	1.2380	1.2630 (2,6)	1.2580	1.2810 (2,7)	1.2660
0.5	1.3	1.2670 (1,2)	1.3030	1.3560 (2,3)	1.3740	1.4100 (2,6)	1.4070	1.4420 (3,6)	1.4250
0.7	1.1	1.3280 (1,2)	1.3790	1.4410 (2,3)	1.4760	1.5100 (3,4)	1.5310	1.5550 (4,5)	1.5650
0.9	0.9	1.3400 (1,2)	1.4060	1.4650 (2,3)	1.5160	1.5420 (3,4)	1.5840	1.5950 (4,5)	1.6310
0.1	0.8	0.7934 (0,3)	0.8009	0.8378 (0,5)	0.8461	0.8599 (0,7)	0.8685	0.8750 (1,8)	0.8809
0.2	0.7	0.8061 (0,3)	0.8337	0.8535 (1,4)	0.8706	0.8752 (1,6)	0.8863	0.8861 (2,7)	0.8936
0.3	0.6	0.8210 (1,2)	0.8553	0.8610 (1,4)	0.8847	0.8828 (2,5)	0.8945	0.8915 (3,6)	0.8981
0.4	0.5	0.8256 (1,2)	0.8659	0.8681 (2,3)	0.8908	0.8859 (3,4)	0.8976	0.8936 (4,5)	0.8994
0.45	0.45	0.8213 (2,1)	0.8671	0.8661 (3,2)	0.8914	0.8850 (4,3)	0.8978	0.8932 (5,4)	0.8994

REFERENCES

[1] Chu, W.W., "Buffer Behavior for Poisson Arrivals and Multiple Synchronous Constant Outputs," IEEE Transactions on Computers **C-19**, No. 6, (June 1970).

[2] Chu, W.W., "Buffer Behavior for Batch Poisson Arrivals and Single Constant Ouput," IEEE Transactions on Communications Technology **Com-18**, No. 5 (October 1970).

[3] Cohen, J.W., *The Single Server Queue* (North Holland, London, 1969).

[4] Cooper, R.B., *Introduction to Queueing Theory* (Macmillan, New York, 1972).

[5] Cooper, R. B., and B. Tilt, "On the Relationship Between the Distribution of Maximal Queue Length in the M/G/1 Queue and the Mean Busy Period in the M/G/1 Queue," Journal of Applied Probability **13**, No. 1 (1976).

[6] Gross, D., and C. M. Harris, *Fundamentals of Queueing Theory*, (Wiley, New York, 1974).

[7] Hoksted, P., "The G/M/m Queue with Finite Waiting Room," Journal of Applied Probability **12**, No. 4 (1975).

[8] Kleinrock, L., *Communicatons Nets: Stochastic Message Flow and Delay* (McGraw Hill, N.Y. 1964; republished by Dover, New York, 1972).

[9] Kleinrock, L., *Queueing Systems*, Volume 1 (Wiley-Interscience, New York, 1975).

[10] Latouche, G., "Exponential Servers Sharing a Finite Storage: Comparison of Space Allocation Policies," Department of Statistics and Computer Science, University of Delaware, Technical Report No. 78/9 (May 1978).

A STREAMLINED SIMPLEX APPROACH TO THE SINGLY CONSTRAINED TRANSPORTATION PROBLEM*

D. Klingman

The University of Texas at Austin
Austin, Texas

R. Russell

The University of Tulsa
Tulsa, Oklahoma

ABSTRACT

A primal simplex procedure is developed to solve transportation problems with an arbitrary additional linear constraint. The approach is a specialization of the Double Reverse Method of Charnes and Cooper. Efficient procedures for pricing-out the basis, determining representations, and implementing the change of basis are presented. These procedures exploit the pure transportation substructure in such a manner that full advantage may be taken of the computational schemes and list structures used to store and update the basis in codifying the MODI method. Furthermore, the pricing-out and change-of-basis procedures are organized in a manner that permits the calculations for one to be utilized in the other. Computational results are presented which indicate that this method is at least 50 times faster than the state-of-the-art LP code, APEX-III. Methods for obtaining basic primal "feasible" starts and "good" feasible integer solutions are also presented.

1. INTRODUCTION

In this paper we address a linear programming problem whose constraints consist of the standard transportation problem plus one extra constraint. Such linear models occur frequently in transportation applications and, as Dantzig [5] points out, the extra constraint often appears in the form of a lower- or upper-bounded partial sum of some of the activities. Charnes and Cooper [2] suggest that in many applications the extra constraint may simply express the proportionality or equality of particular variables for example, forcing two routes to have equal flow. The warehouse funds-flow model discussed by Charnes and Cooper [2] is a specific application which may appear in the form of a transportation problem with one extra constraint. Another application is a constrained version of Wagner's employment-scheduling problem [20].

*This research was partly supported by ONR Contracts N00014-67-A-0126-0008 and N00014-67-A-0126-0009 with the Center for Cybernetic Studies, The University of Texas at Austin, and by ONR Contract N00014-76-C-0383 with Decision Analysis and Research Institute. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Operations research literature contains a number of ingenious techniques for transforming a transportation problem with an extra linear constraint into a larger equivalent transportation problem. For instance, [2,3,4,5,13,15,17,18] are indicative of the interest in this problem. However, these transformations are not possible with an arbitrary extra constraint. Thus, in this paper we develop a primal simplex-solution procedure which exploits fully the topological structure embedded in this problem. One of the main purposes is to specify computationally efficient procedures for determining the dual evaluators and updating the basis. Additionally, an efficient method for finding the representation of any entering vector with respect to the current basis is developed. Thus our solution method, like the MODI method [5] and the Row-Column Sum Method [2], only requires the knowledge of the basis rather than the basis inverse to be able to implement the simplex method efficiently. Furthermore, these procedures exploit the structure in such a manner that they allow us to take full advantage of the computational schemes and list structures [1,7,9,12] used to store and update the basis in codifying the Row-Column Sum Method. Our approach also allows for the use of advanced start procedures. For instance, we show how the Row Minimum Rule [6] and Vogel's Approximation Method [17] for the pure transportation problem can be extended to yield good starts for the singly constrained problem.

Since the extra constraint may destroy the integrality property of the pure transporation problem, an optimal solution may be noninteger. Thus another purpose of this paper is to indicate an efficient way to find a "good" integer solution to the singly constrained transportation problem.

Our solution procedure for the singly constrained transportation problem may be viewed as a specialization of the Double Reverse Method of Charnes and Cooper [2]. Further, our approach uses this pedagogical overview to exploit the subproblem structure of the singly constrained transporation problem by viewing the problem in a graphical context. In particular, the basis partitioning simplex algorithm is adapted to solve this class of problems by storing the coefficient matrix and the basis matrix as graphs using computer list structures.

The use of such structures reduces both the amount of work required to perform the basic simplex steps and the amount of computer memory required to store essential data. The matrix operations of finding the representation of an entering vector and determining updated dual variable values are performed by tracing paths within the basis graph. Since the graphs contain only the nonzero entries in the problem (basis) the list procedures [1,7,9,12] eliminate checking or performing unnecessary arithmetic operations on zero elements. In addition, by exploiting the near-triangular basis properties of such problems, this approach allows the basis inverse to be stored implicitly as a graph. This graph is updated during basis exchange steps by simply changing a few pointers in the list structures. This entirely eliminates the arithmetic operations normally used to update the basis inverse and also eliminates round-off error.

The advantages of this approach are dramatically illustrated by the computational results in section 7. These results indicate that this approach is at least 50 times faster than the state-of-the-art linear programing techniques contained in APEX-III.

2. TOPOLOGICAL PROPERTIES OF THE SINGLY CONSTRAINED TRANSPORTATION PROBLEM

The singly constrained transportation problem may be stated as

$$\text{minimize} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} + 0S$$

$$\text{subject to} \sum_{j=1}^{n} x_{ij} = a_i,\ i = 1, 2, \ldots, m,\ a_i \geqslant 0$$

(1)
$$\sum_{i=1}^{m} x_{ij} = b_j,\ j = 1, 2, \ldots, n,\ b_j \geqslant 0,$$

$$x_{ij} \geqslant 0 \text{ for all } i \text{ and } j,$$

(2)
$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} x_{ij} - S = k \text{ (where some } f_{ij} \text{ may be zero)},$$

$$S \geqslant 0,$$

$$\text{as we assume } \sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j.$$

The dual of this problem is

$$\text{maximize} \sum_{i=1}^{m} a_i R_i + \sum_{j=1}^{n} b_j K_j + k\delta$$

(3)
$$\text{subject to } R_i + K_j + f_{ij}\,\delta \leqslant c_{ij},\ i = 1, 2, \ldots, m,$$

$$\delta \geqslant 0,$$

$$R_i,\ K_j - \text{unrestricted for all } i \text{ and } j.$$

Graphically, the problem may be portrayed as a bipartite graph whose arcs have "flags" corresponding to the coefficients in constraint (2). In particular, the arc incidence conditions, as defined by (1), define a transporation network and the extra constraint (2) is simply viewed as placing joint restrictions over the arcs. Because of constraint (2), certain columns of the coefficient matrix will have three nonzero entries; the arcs corresponding to these columns are called the arcs with flags, or "flagged" arcs.

One of the columns of the coefficient matrix has only one entry. This column corresponds to the slack variable S which appears in (2). The slack variable is another special type of arc which may be viewed as being attached to any node. For convenience, we assume that it is attached to node 1. In the literature, a slack arc has been referred to as a slack loop (see Dantzig [5]). The loop structure inherent in the singly constrained transportation problem will play an important role in the subsequent computational development. Figure 1 illustrates the complete graph structure of a 2×3 singly constrained transporation problem. The number appearing in the rectangle on each arc corresponds to the cost while the number appearing in the semicircle corresponds to the flag value associated with the arc.

A basis for the singly constrained transportation problem has a structure which is closely related to the pure transportation problem. It is generally known that the basis of a pure transportation problem is a spanning tree [5], and the basis has rank $m + n - 1$. Since the extra constraint (2) involves the slack variable S, the singly constrained transportation problem will clearly have a basis of rank $m + n$. The following remark relates the basis structure of the pure and singly constrained problems; it also leads to a complete characterization of the basis.

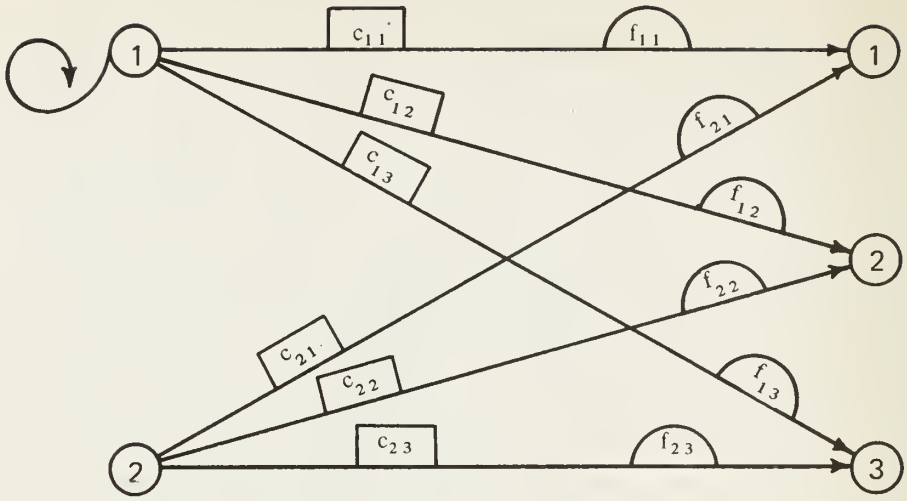

FIGURE 1. Illustration of graph structure of 2×3 singly constrained problem.

REMARK 1: Consider a basis of $m + n$ columns (arcs) for the singly constrained transporation problem. Delete from these columns (arcs) the component corresponding to constraint (2). From the set of $m + n$ reduced columns (arcs), there exists $m + n - 1$ columns (arcs) which form a basis for the underlying pure transportation problem.

PROOF: Let B be any basis for the singly constrained transportation problem. B is an $(m + n + 1) \times (m + n)$ matrix. Delete the row of B corresponding to the components associated with the extra constraint (2); the rank of the resulting matrix B' is at least $m + n - 1$ since the row rank of B is $m + n$. Furthermore, since the rank of the coefficient matrix of the pure transportation problem is $m + n - 1$, the row rank of B' is at most $m + n - 1$. Thus the rank of B' is $m + n - 1$. Hence there exists $m + n - 1$ independent arcs among the $m + n$ arcs in B'. By definition these $m + n - 1$ arcs form a spanning tree for the pure transportation problem.

It is a fundamental property of spanning trees that a spanning tree plus one additional arc incident on the tree nodes contains precisely one loop [14,18]. The basis for the singly constrained transporation problem can assume one of two possible loop structures. In one case the basis contains $m + n$ arcs which are indigenous to the underlying pure transporation problem. In this case the basis can be viewed as a tree with an extra arc. In the other case, the basis contains the slack arc and $m + n - 1$ arcs which are indigenous to the pure transportation problem. In this case the basis can be viewed as a slack loop with an attached tree.

In the next sections an efficient primal simplex algorithm is developed for the singly constrained transportation problem. The efficiency of this algorithm is derived from exploiting the near-triangular structure of the basis in a graphical context. Specifically, we view the basis as being stored as a rooted tree plus an extra arc using the efficient list structures and labeling procedures [1,7,9,12] for representing a spanning tree as a rooted tree. These list structures and labeling techniques provide the means for performing the two types of tree traversal needed to execute the basic simplex steps.

Specifically, these lists allow one to find efficiently the unique path in a spanning tree joining one node to another node. This feature allows one to find easily the loop (called the basis

loop) created in the spanning tree by the extra basis arc. Similarly, the loop created in the spanning tree by adding an entering nonbasic arc is called a nonbasis loop. In section 4, we show that the representation of any nonbasic arc in terms of the basis can be easily determined by reference to the basis loop arcs and nonbasis loop. Additionally, these lists can be used to locate all nodes situated below a given node in the rooted tree. (Note that this discussion assumes that the root is the highest node in the tree.) This feature is used to update the dual variable values quickly and efficiently.

Dual variable values are updated in some cases by locating all nodes situated "below" a given node of the arc leaving the basis. In other cases, when all the dual variables values change, this may be efficiently done by starting at the root of the tree and visiting each node of the tree. The efficiency of this approach derives from the facts that only nonzero coefficients are stored, that operations involving a basis matrix inverse may be performed using the basis graph, and that basis exchange steps involve only the updating of list structure pointers.

3. EFFICIENT IMPLEMENTATION OF THE PRICING-OUT PROCEDURE

In this section computationally efficient procedures for determining the dual evaluator values, and thus the updated costs, are presented. Pricing out the basis (that is, determining the dual evaluator values) is equivalent to finding the simultaneous solution of the equality form of the dual constraints in (3) associated with the primal variables in the basis. The resulting set of dual equations forms an $(m + n) \times (m + n + 1)$ system of equations. Our procedure for determining the dual evaluators greatly simplifies the solution of the $(m + n) \times (m + n + 1)$ system of dual equations. The procedure is a two-step approach in which the value of the dual evaluator, δ, is initially obtained and the values of the remaining dual evaluators are subsequently determined.

In the previous section, we noted that there are two types of basis graph structures. In one case the graph contains a loop consisting of a single slack arc; in the other case the loop consists of a sequence of an even number of arcs.

For a basis graph which includes a slack (artificial) arc, the Complementary Slackness Theorem implies that the dual evaluator, δ, associated with the extra constraint, can be immediately priced to zero (M). Since the remaining basic arcs form a spanning tree, the remaining dual evaluators may be efficiently determined if the basis has been stored as a rooted tree using the labeling in [1,7,9,12]. Thus, one can start at the root of the tree, setting its dual variable to zero, and sequentially visit all nodes in the tree using the dual equations to calculate their values.

On the other hand, a graph which contains a "nonslack" basis loop requires a more elaborate treatment. (For editorial convenience we shall denote such a loop by $[(1,1),(2,1),(2,2),(3,2),\ldots,(p,p),(1,p)]$. Figure 2 illustrates this loop with associated costs c_{ij}, flag values f_{ij}, and dual evaluators δ, R_i, and K_j for each appropriate node.) This case may be handled by first determining the value of the dual evaluator δ and then pricing out the rooted tree in a manner analogous to the slack-loop case to find the R_i and K_j dual evaluators.

Thus, maintaining a partition of the basis as a rooted spanning tree and one other arc enables us to determine immediately the remaining dual evaluators once the value of δ has been determined. Further, we show that the dual evaluator δ can be easily determined by considering only the dual equations associated with basic variables in the basis loop.

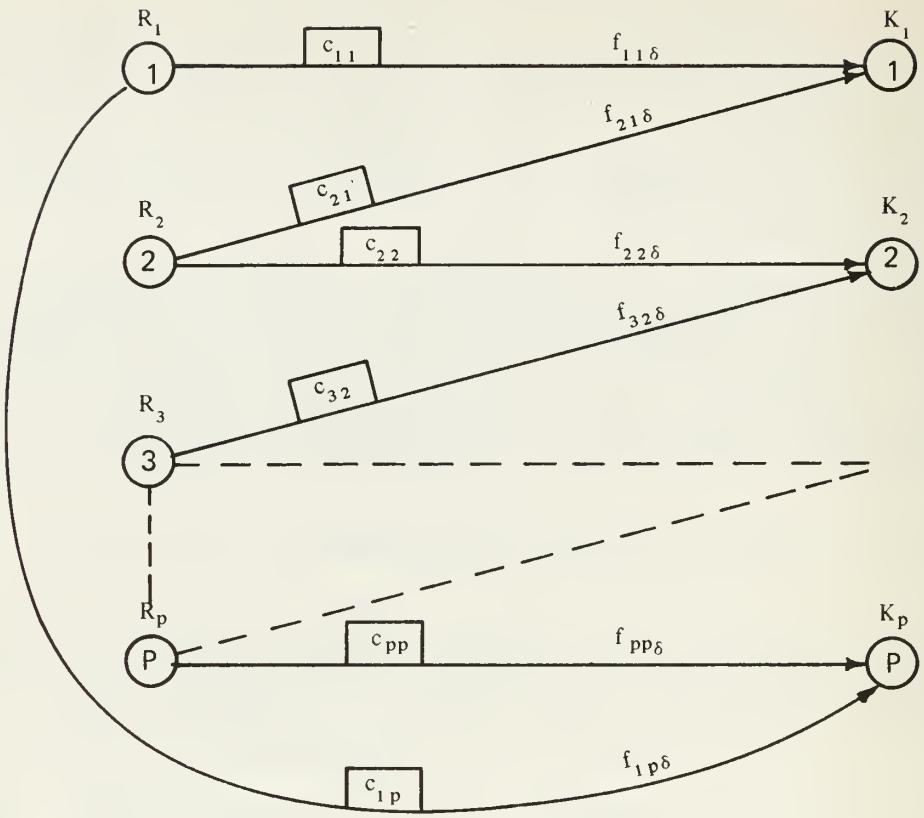

FIGURE 2. A nonslack loop with associated dual evaluators.

Since the dual constraints associated with the basic arcs are binding at optimality, the dual evaluators R_i, K_j and δ associated with the loop in Figure 2 may be found by solving the system of simultaneous equations;

$$
\begin{aligned}
(4)\qquad & R_1 + K_1 && + f_{11}\,\delta = c_{11}\\
& K_1 + R_2 && + f_{21}\,\delta = c_{21}\\
& R_2 + K_2 && + f_{22}\,\delta = c_{22}\\
& K_2 + R_3 && + f_{32}\,\delta = c_{32}\\
& \cdot && \cdot \quad\ \cdot\\
& \cdot && \cdot \quad\ \cdot\\
& \cdot && \cdot \quad\ \cdot\\
& && R_p + K_p + f^{pp}\,\delta = c_{pp}\\
& R_1 && + K_p + f_{1p}\,\delta = c_{1p}.
\end{aligned}
$$

THEOREM 1: An explicit solution value for δ in (4) is

$$
(5)\qquad \delta = \frac{\sum_{i=1}^{p} c_{ii} - \sum_{i=2}^{p} c_{ii} - 1 - c_{1p}}{\sum_{i=1}^{p} f_{ii} - \sum_{i=2}^{p} f_{ii} - 1 - f_{1p}}
$$

PROOF: The entire system of dual equations associated with the basis has one degree of freedom. Thus it will be assumed, with loss of generality, that $R_1 = 0$. The proof involves solving the first equation in (4) for K_1 in terms of δ and then performing the obvious substitution into the next equation in (4), solving repeatedly for each successive variable in terms of the preceding, and hence in terms of δ.

Solving the first equation in (4) for K_1 in terms of δ, we obtain

$$K_1 = c_{11} - f_{11}\,\delta. \tag{6}$$

Substituting (6) into the second equation in (4), we have

$$R_2 = (c_{21} - f_{21}\,\delta) - (c_{11} - f_{11}\,\delta). \tag{7}$$

Using (7) in the third equation yields

$$K_2 = (c_{22} - f_{22}\,\delta) + (c_{11} - f_{11}\,\delta) - (c_{21} - f_{21}\,\delta), \tag{8}$$

which secures

$$\begin{aligned} R_3 = {} & (c_{32} - f_{32}\,\delta) + (c_{21} - f_{21}\,\delta) \\ & - (c_{22} - f_{22}\,\delta) - (c_{11} - f_{11}\,\delta). \end{aligned} \tag{9}$$

Thus, repeated performance of these operations yields,

$$R_p = \sum_{i=2}^{p} (c_{i-1} - f_{i-1}\,\delta) - \sum_{i=2}^{p} (c_{i} - f_{i}\,\delta), \tag{10}$$

$$K_p = \sum_{i=1}^{p} (c_{i} - f_{i}\,\delta) - \sum_{i=2}^{p} (c_{i-1} - f_{i-1}\,\delta). \tag{11}$$

Focusing attention on the last equation of (4) yields

$$f_{1p}\,\delta = \left(\sum_{i=1}^{p} f_{i} - \sum_{i=2}^{p} f_{i-1}\right)\delta + \sum_{i=2}^{p} c_{i-1} + c_{1p} - \sum_{i=1}^{p} c_{i},$$

$$\left(\sum_{i=1}^{p} f_{i} - \sum_{i=2}^{p} f_{i-1} - f_{1p}\right)\delta = \sum_{i=1}^{p} c_{i} - \sum_{i=2}^{P} c_{i-1} - c_{1p}. \tag{12}$$

Assuming that $\sum_{i=1}^{p} f_{i} - \sum_{i=2}^{p} f_{i-1} - f_{1p}$ is not equal to zero, expression (12) yields the desired result. By focusing attention on the loop portion of the basis, we may show that this assumption is valid. In particular, we form an "alternating" linear combination of the columns associated with the loops where, in traversing the loop clockwise, the loop arcs traversed in the forward direction are multiplied by $+1$ and those traversed in the backward direction are multiplied by -1. The resulting linear combination is

$$+1\hat{P}_{11} - 1\hat{P}_{21} + 1\hat{P}_{22} - 1\hat{P}_{32} + \ldots - \hat{P}_{1p} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \sum_{i=1}^{p} f_{i} - \sum_{i=2}^{p} f_{i-1} - f_{1p} \end{pmatrix} \tag{13}$$

where $\hat{P}_{ij}$ denotes that column of the coefficient matrix associated with variable x_{ij}. Thus, if $\sum_{i=1}^{p} f_{ii} - \sum_{i=2}^{p} f_{ii-1} - f_{1p}$ is zero, then the loop arcs would be linearly dependent, a contradiction, thus completing the proof.

The graphical technique proposed by the theorem for calculating the value of the dual variable δ may be described in operational terms as follows. Keep the partitioned basis stored as a spanning tree and an extra arc. Starting at origin loop node 1, set dual variable $R_1 = 0$. Next, traverse the basis loop in a clockwise direction. While traversing the basis loop keep a sum C^+ and a sum F^+ of the cost coefficients and flag values, respectively, of the arcs traversed in a forward direction. Likewise, form a sum C^- and a sum F^- of the cost coefficients and flag values, respectively, of the arcs traversed in a backward direction. Then $(C^+ - C^-) / (F^+ - F^-)$ is the desired value of δ.

From a count of operations it is not difficult to see that the foregoing technique constitutes the most efficient means for computing δ via expression (5). After calculating the value δ, start at the root of the tree, setting its dual evaluator to zero. Whereupon, the other dual evaluators are immediately determined in a cascading fashion if we move down the tree and identify the value for each R_i or K_j from its predecessor by using the equation $R_i + K_j + \delta f_{ij} = c_{ij}$. Highly efficient labeling procedures for traversing the tree to initialize and update these dual evaluators are described in [1,7,9,12].

4. CHANGE OF BASIS AND UPDATING PROCEDURES

In this section we present computationally efficient procedures for implementing the change of basis procedure and for updating the dual evaluators as well as the basic flows. The discussion provides a connection between the calculation of δ in the preceeding section and calculations involved in the change of basis, enabling the latter to be carried out with marginal additional effort.

In order to execute a change of basis it is necessary to find the representation of a nonbasic arc (r, s) in terms of the basic arcs. Because of the structure and previous calculations, the determination of the representation is relatively simple. We have three cases:

CASE 1: Suppose the arc (r, s) is the slack arc. (Let us denote the column of the coefficient matrix associated with arc (r, s) as $\hat{P}_{rs}$ and let P_{ij} denote that portion of $\hat{P}_{ij}$ associated with the underlying pure transportation problem; i.e., $\hat{P}_{ij}$ is an $m + n + 1$ component vector (since it contains a flag component) whereas P_{ij} has only $m + n$ components.) Prior to the entrance of the slack arc in the basis, the basis loop must consist of a "nonslack" loop. Then from (13) we can find the representation of the slack arc by solving for θ such that

$$(14) \quad \theta \begin{pmatrix} 0 \\ \vdots \\ 0 \\ F^+ - F^- \end{pmatrix} = -e^{m+n+1} \text{ (where } e^{m+n+1} \text{is the } m+n+1 \text{ column of an identity matrix).}$$

Clearly,

(15) $\theta = -1 / (F^+ - F-)$ (Note that we are assured θ does not equal zero.)

Thus the slack arc is represented solely in terms of the basis loop arcs and this representation is given by

(16) $$+\theta\hat{P}_{11} - \theta\hat{P}_{21} + \theta\hat{P}_{22} - \ldots - \theta\hat{P}_{lp} = -e^{m+n+1}$$

Note that no additional calculations are required, since $F^+ - F^-$ was determined in the pricing-out procedure of the preceding section. Further note that the coefficients in the representation are equal in magnitude. Thus to determine the column vector (arc) to leave the basis only requires checking the sign of θ and then finding the minimum flow value associated with the loop arcs traversed in a forward (or backward) direction if θ is positive (or negative). Thus no ratios have to be computed to find the vector to leave the basis. Also, the flow values associated with the basis can be immediately updated with a minimum of computational effort by simply adding or subtracting the current flow value of the arc leaving the basis from the appropriate basis loop arcs and setting the flow on the entering slack arc to this value divided by θ. All other basic flows remain unchanged.

Since the entrance of a slack arc in the basis causes both the value of δ and basis loop structure to change, the dual evaluators associated with the new basis cannot be conveniently updated; thus, the dual evaluators would have to be calculated as in section 2.

CASE 2: Now, suppose that the incoming arc (r,s) is not a slack arc, but that the slack arc is contained in the basis. By Remark 1, the current basis may be partitioned into a spanning tree and the slack arc. Focusing only on the spanning tree portion of the basis, we can determine the unique undirected path which leads from origin node r to destination node s. Such a path may be denoted $\{(r,j_1), (i_1,j_1), (i_1, j_2), \ldots (i_k,s)\}$. (Note this path will contain an odd number of arcs.) We make the following observation:

(17) $$+1\hat{P}_{rj_1} - 1\hat{P}_{i_1j_1} + 1\hat{P}_{i_1j_2} - \ldots + 1\hat{P}_{i_ks} = \begin{pmatrix} P_{rs} \\ f_{rj_1} - f_{i_1j_1} + f_{i_1j_2} - \ldots + f_{i_ks} \end{pmatrix},$$

i.e., the linear combination formed by multiplying arcs in the path which are traversed in the forward direction by $+1$ and arcs in the path which are traversed in the backward direction by -1 yields a vector which equals $\hat{P}_{rs}$ is all components, except possibly the $m + n + 1$st. Thus we use the slack basis loop arc and the results of (17) to form the following equation:

(18) $$\begin{pmatrix} P_{rs} \\ f_{rj_1} - f_{i_1j_1} + f_{i_1j_2} - \ldots + f_{i_ks} \end{pmatrix} + \theta \begin{pmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} P_{rs} \\ f_{rs} \end{pmatrix} = \hat{P}_{rs}.$$

Solving yields

(19) $$\theta = -f_{rs} + f_{rj_1} - f_{i_1j_1} + f_{i_1j_2} - \ldots + f_{i_ks}.$$

Thus, with the value of θ as computed in (19), we have representation of arc (r, s) as

(20) $$+1\hat{P}_{rj_1} - 1\hat{P}_{i_1j_1} + 1\hat{P}_{i_1j_2} + \ldots + 1\hat{P}_{i_ks} + \theta\hat{P}_S = \hat{P}_{rs}$$

where $\hat{P}_S$ is the column vector corresponding to the slack variable S.

The determination of the column vector (arc) to leave the basis is again amenable to computational simplification; we have two cases. If θ in (20) is less than or equal to zero, the computation is particularly simple. The determination of the column vector (arc) to leave the basis only requires finding the minimum flow value Y_p associated with the unique path arcs which are traversed in a forward direction. The arc associated with the minimum flow value Y_p is the arc to leave the basis. The basic flows may be updated by assigning a flow of Y_p to the incoming arc, adding Y_p to the flow on the "backward arcs," subtracting Y_p from the flow on the "forward arcs" in the unique path, and adding $-\theta Y_p$ to the flow on the slack arc.

If θ in (20) is greater than zero, we need to compute two values. Let Y_p be calculated as above, and let Y_S equal the flow on the slack arc divided by θ. Then the arc to leave the basis will be the arc whose flow value is equal to the minimum of Y_p and Y_S. If the arc to leave the basis is in the unique path, then the flows may be updated as above (i.e., as in the previous case when $\theta \geqslant 0$). If the outgoing arc is the slack arc, then we add and subtract Y_S from the appropriate arcs in the unique path. All other basic flows remain unchanged.

In general, when the arc to leave the basis is not contained in the basis loop, then it is computationally advantageous to update the dual evaluators rather than to calculate them directly as in section 2. Thus when the incoming arc does not replace the slack arc, we will update the dual evaluators. In this case the origin node k and the destination node t of the outgoing arc (k, t) will be contained in the spanning-tree portion of the basis. Deleting the arc (k, t) from the basis splits the spanning tree into two disjoint subtrees, say T_1 and T_2. One of these, say T_1, is attached to the loop and T_2 is disjoint from the remainder of the basis. All dual evaluators associated with nodes not in T_2 (including δ) will remain unchanged. To update the dual evaluators associated with nodes in T_2, we compute

$$\bar{c} = R_r + K_s + \delta f_{rs} - c_{rs}, \tag{21}$$

which is the current updated cost associated with the incoming arc (r,s). The dual evaluators in T_2 may be updated in the following manner. If $r \in T_2$, subtract $\bar{c}$ from R_r and all other R_k such that $k \in T_2$. Further, add $\bar{c}$ to all dual evaluators K_j such that $j \in T_2$. If $s \in T_2$, subtract $\bar{c}$ from K_s and all other K_j such that $j \in T_2$. Further, add $\bar{c}$ to all dual evaluators R_i such that $i \in T_2$. A proof of the validity of the preceding updating procedure follows straightforwardly from the facts established in the papers [1,7,9,12].

CASE 3: We finally consider the case in which the arc (r,s) to enter the basis is not a slack arc and a slack arc is not contained in the current basis. Again we partition the basis into a spanning tree plus one additional arc. Using the results of (13) and (17) we obtain

$$(f_{ri_1} - f_{i_1 j_1} + \ldots + f_{i_k s}) + \theta \begin{pmatrix} 0 \\ \vdots \\ 0 \\ F^+ - F^- \end{pmatrix} = \begin{pmatrix} P_{rs} \\ f_{rs} \end{pmatrix} = \hat{P}_{rs} \tag{22}$$

Solving yields

$$\theta = \frac{f_{rs}}{F^+ - F^-} - \frac{[f_{ri_1} - f_{i_1 j_1} + \ldots + f_{i_k s}]}{F^+ - F^-} \tag{23}$$

Thus,

$$+ 1\hat{P}_{ri_1} - 1\hat{P}_{i_1 j_1} + 1\hat{P}_{i_1 j_2} - \ldots + 1\hat{P}_{i_k s} + \theta\hat{P}_{11} - \theta\hat{P}_{21} + \theta\hat{P}_{22} - \ldots - \theta\hat{P}_{1n} = \hat{P}_{rs}, \tag{24}$$

and the "representation" of $\hat{P}_{rs}$ is given in (24). Again little additional computation is required, since $F^+ - F^-$ has previously been determined, and computationally efficient methods of determining the unique path are available [1,7,9,12].

In determining the column vector (arc) to leave the basis and subsequent flow changes, we first note that some of the column vectors (arcs) in (24) may appear twice. That is, some of the column vectors (arcs) in the unique path may also be contained in the basis loop. When the column vectors (arcs) in (24) appear only once, certain computational simplifications arise. In this case, we have two subcases. First, if $\theta = 0$ we have basically the same situation as in case 2 when $\theta \leqslant 0$. Thus the change of basis procedure would be performed in a manner analogous to case 2. Also, calculate Y_l^+, the minimum flow associated with a "forward loop arc," and Y_l^-, the minimum flow associated with a "backward loop arc." If $\theta > 0$ $(\theta < 0)$, compare Y_p with Y_l^+ / θ $(-Y_l^- / \theta)$. If Y_p is the smaller of the two values, then the vector (arc) to leave the basis is in the unique path. The flows may be updated by assigning a flow of Y_p to the incoming arc (r, s), adding or subtracting Y_p from the appropriate unique path arcs, and adding or subtracting $\theta\ Y_p$ from the appropriate loop arcs. (Notice that in this case Y_l^+ and Y_l^- may be easily updated.) If Y_p is the larger of the two values, then the vector (arc) to leave the basis is in the loop. When $\theta > 0$ $(\theta < 0)$ the incoming arc (r, s) is assigned a flow of Y_l^+ / θ $(-Y_l^- / \theta)$, and Y_l^+ $(-Y_l^-)$ is added to the appropriate loop arcs. Similarly, Y_l^+ / θ $(-Y_l^- / \theta)$ is added to the appropriate arcs in the unique path. All other basic flows remain unchanged.

When the vectors (arcs) in the unique path are contained in the loop portion of the basis, no simplifications arise. (Note that in this case $\theta \neq 0$.) In this case the "usual" linear programming pivoting process must be used to determine the vector (arc) to leave the basis and the updated basic flows.

Again, we may consider updating the dual evaluators. If the vector (arc) to leave the basis is in the loop, then updating is not computationally advantageous and the direct method of calculating the dual evaluators should be applied. However, if the vector (arc) leaving the basis is not in the loop portion of the basis we may update the dual evaluators in a manner completely analogous to the updating in case 2.

The previous procedures for determining representations and updating basic flows and dual evaluators have been designed to minimize computational effort. However, another important advantage of these procedures involves the accumulated round-off error. The pricing-out procedure minimizes round-off error in that the calculation of δ contains all of the round-off error; all other dual evaluators (whether calculated directly or by updating) are derived from original parameters by addition and subtraction. Likewise, the calculation of the flow values minimizes round-off error. However, in successive pivoting some error may accumulate, and it may be desirable to generate an exact representation of the original stipulating vector.

5. OBTAINING A BASIC PRIMAL FEASIBLE SOLUTION

We now consider the problem of obtaining a basic "feasible" start for the singly constrained transportation problem. The problem is easily handled by extending the primal start algorithms which exist for the pure transportation problem. A variety of start algorithms exists for obtaining a basic feasible start for the pure transportation problem. Primal start methods include the Northwest Corner Rule [5], Vogel Approximation Method (VAM) [17], Row

Minimum Rule [6], Row Column Minimum Rule [5], Modified Row Minimum Rule [8], Matrix Minimum Rule [13], and Matrix Minimum Rule by Ranking Method [16]. Another way of obtaining a good basic start is to use an optimal solution to the underlying pure transportation problem.

REMARK 2: A basic "feasible" solution for the singly constrained transportation problem may be obtained by applying one of the above starting methods to the underlying pure transportation problem and then adding an appropriate slack or artificial (as determined by (2)) to the pure transportation starting basis.

PROOF: Since the pure transportation start yields a feasible spanning tree for constraints (1) of the singly constrained problem, we need only augment the pure starting basis by an extra variable (slack or artificial) whose value will satisfy (2). Clearly the augmented vectors are linearly independent and satisfy (1) and (2). Thus a basic solution is easily determined which satisfies primal feasibility, possibly in an artificial sense.

6. OBTAINING A FEASIBLE INTEGER SOLUTION

An optimum solution to a singly constrained transportation problem (as with most linear programming problems) may be noninteger. In this section we present a simple means of obtaining a "good" feasible integer solution to the singly constrained transportation problem. (Throughout the section we assume that the transporation parameters are integer, that is, we assume that the cost coefficients and supply and demand amounts are integer.) Such an integer solution may serve as an end in itself or provide an upper bound on the optimum integer solution. If an optimum integer solution is desired, then a relatively tight upper bound should provide a useful tool for implementing a branch-and-bound solution procedure. Furthermore, the feasible integer solution may be obtained with little additional computation beyond the nonintegral optimum.

REMARK 3: A feasible integer solution to the singly constrained transportation problem may be obtained by pivoting the slack variable (in constraint (2)) into the optimal basis.

PROOF: Of course, if the slack variable is in the optimal basis, then the optimal solution is integer, and no additional work is necessary. Thus, it is only necessary to show that it is always possible to pivot the slack variable into the optimal basis and maintain primal feasibility. To do this, it is necessary to show that the representation of the slack variable contains at least one positive component. From expression (17) we see that half of the coefficients in its representation have a value of θ and the other half a value of $-\theta$. Furthermore, expression (16) guarantees that θ does not equal zero. Hence it will always be possible to pivot the slack variable into the basis and maintain primal feasibility.

7. COMPUTATIONAL RESULTS

A computer code was developed for solving singly constrained transportation problems in order to confirm the relative efficiency of the preceding algorithms. The code (TRANSC) uses the threaded index method [12] for storing the basis in graph form and also uses the procedures presented earlier. The code is written entirely in standard FORTRAN IV and was initially debugged and tested using the RUN compiler on a CDC 6600 computer. Under these conditions it occupies a total of $6(M + N) + 3A + 6600$ words of central memory, where M and N denote the number of origins and destinations, respectively, and A the number of arcs.

Following the development of TRANSC a computational study was conducted to determine good start and pivot rules to use in conjunction with the preceding algorithm and to evaluate the adequacy of the integer solution provided by pivoting the slack variable into the basis. To conduct the testing, 35 feasible problems were randomly generated. The test results indicated that the best start rule was to "introduce" the extra constraint to the optimal basis for the underlying transportation problem and to use a multipricing (candidate list) pivot criterion of size 8-16. A complete description of a T-R candidate list follows:

> Examine sequentially each of the origins for the underlying transportation problem. For each origin, select the nonbasic arc that violates dual feasibility by the largest amount. Add this arc (if one exists) to the list and proceed to the next origin. This is done until the list contains R entries or until all of the origins have been explored. From the list the "best" arc is selected to enter the basis. The list is accessed after each pivot until it is void of eligible edges or until it has been accessed T times. At such time, the list is refilled by again examining each origin, starting at the last examined origin. When the last origin is encountered, the dual constraint of the slack variable S is checked before proceeding to the first origin. If a complete pass through the origins is made without filling the list, the size of R is reset to the number found and T is reduced to $1/2R$.

This type of procedure has been found to be exceptionally efficient for pure and generalized network codes. For this reason, the effect of the initial values of T and R for the singly constrained code were tested.

In general, our testing indicates that a singly constrained transportation problem requires approximately three times as much time to solve as the underlying transportation problem. The number of nonzero coefficients in the extra contraint affects the number of times the dual variable δ changes. Consequently, solution time tends to increase as the number of nonzero coefficients increases. Further, the integer solution obtained by pivoting the slack into the optimal continuous basis never increased the objective function value by more than 5% and often by only 1%. Thus, a good integer solution was always obtained.

The next phase of testing was to compare the best version of singly constrained code TRANSC with a widely used standard linear programming package, APEX III. APEX III is distributed by Control Data Corporation, and the tests were conducted on a CDC CYBER-74 computer. In this case, the three problems indicated in Table 1 were used as test data. These problem range from a 50 × 50 constructed transportation problem to a 500 × 500 node constrained transportation problem. The basis of comparison, however, is not CPU seconds, but rather a CDC accounting measure called an SBU. The total SBU count for a job is an accumulation of CPU seconds used, I/O functions performed, and central-memory requirements. This measure seemed appropriate, especially since it allowed the comparison of actual dollar amounts. CDC charges customers a minimum of $0.18 per SBU used.

The results of the comparison of TRANSC and APEX III are contained in Table 2. In this case, TRANSC was from 45.3 times more efficient (on a small-sized problem) to 67.96 times more efficient (on a large scale problem).

Table 1. *Problem Specifications*

Problem	Origins	Destinations	Arcs	Cost Range		Total Supply
				Min	Max	
1	50	50	500	1	100	100000
2	200	700	5000	1	100	100000
3	500	500	5000	1	100	100000

Extra Constraint				
Problem	Nonzero Coefficients	Coefficient Range		Right-Hand Side Value
		Min	Max	
1	3000	−1.5	1.5	50000
2	4000	−2.0	2.0	20000
3	6000	−1.0	1.0	10000

Table 2. *TRANSC vs APEX-III*

Problem	TRANSC		APEX-III	
	SBU's[1]	Cost[2]	SBU's[1]	Cost[2]
1	0.72	0.13	32.77	5.89
2	13.83	2.49	925.18	166.53
3	17.26	3.10	1170.42	210.68[3]

[1]System Billing Unit
[2]Computed at $0.18 per SBU
[3]Terminated after 10,000 iterations
Objective function value = 2.5 billion
Optimal value = 6 million

BIBLIOGRAPHY

[1] Barr, R. F. Glover, and D. Klingman, "Enhancements of Spanning Tree Labeling Procedures for Network Optimization," Research Report CCS 262, Center for Cybernetic Studies, The University of Texas at Austin, (1976).

[2] Charnes, A., and W. W. Cooper, *Management Models and Industrial Applications of Linear Programming*, Vols. I and II (Wiley, New York, 1961).

[3] Charnes, A. F. Glover, and D. Klingman, "The Lower Bounded and Partial Upper Bounded Distribution Model," Naval Research Logistics Quarterly *18* (2), 277-281, (1971).

[4] Charnes A, and D. Klingman, "The Distribution Problem with Upper and Lower Bounds on the Node Requirements," Management Science *16*, (9), 638-642, (1970).

[5] Dantzig, G. G., *Linear Programming and Extensions* (Princeton University Press, Princeton, New Jersey, 1963).

[6] Dennis, J. B., "A high-Speed Computer Technique for the Transportation Problem," Journal of the Association for Computing Machinery *8*, 132-153, (1958).

[7] Glover, F., D. Karney, and D. Klingman, "The Augmented Predecessor Index Method for Locating Stepping-Stone Paths and Assigning Dual Prices in Distribution Problems," Transportation Science *6*, (2), 171-179, (1972).

[8] Glover, F., D. Karney, D. Klingman, and A. Napier, "A Comparison of Computation Times for Various Starting Procedures, Basis Change Criteria, and Solution Algorithms for Distribution Problems," Management Science *20*, (5), 793-813, (1974).

[9] Glover, F., and D. Klingman, "Locating Stepping-Stone Paths in Distribution Problems via the Predecessor Index Method," Transportation Science, *4* (2), 220-225, (1970).

[10] Glover, F. and D. Klingman, "A Note on Computational Simplifications in solving Generalized Transportation Problems," Transportation Science, *7* (4), (1973).

[11] Glover, F., D. Klingman, and G. T. Ross, "On Transforming Constrained Transportation Problems into Pure Transportation Problems," Naval Research Logistics Quarterly *21* (2), 247-253, (1974).

[12] Glover, F., D. Klingman, and J. Stutz, "The Augmented Threaded Index Method for Network Optimization," INFOR *12* (3), 293-298, (1974).

[13] Hadley, G., *Linear Programming* (Addison-Wesley, Reading, Massachusetts, 1962).

[14] Johnson, E., "Programming in Networks and Graphs," ORC 65-1, University of California, Berkeley, (1965).

[15] Klingman, D., "Finding Equivalent Network Formulation for Constrained Network Problems," *Management Science 23*, (7), 737-745, (1977).

[16] Lee, S., "An Experimental Study of the Transportation Algorithms," Masters Thesis, Graduate School of Business, University of California at Los Angles, (1968).

[17] Reinfeld, N. V., and W. R. Vogel, *Mathematical Programming*, (Prentice-Hall, Englewood Cliffs, New Jersey, 1958).

[18] Simmonard, M., *Linear Programming*, translated by W. S. Jewell (Prentice-Hall Englewood cliffs, New Jersey, 1966).

[19] Srinivasan, V., and G. L. Thompson, "Accelerated Algorithms for Labeling and Relabeling of Trees, with Application to Distribution Problems," Journal of the Association for Computing Machinery.

[20] Wagner, H. M., *Principles of Operations Research,* (Prentice-Hall, Englewood Cliffs, New Jersey, 1969).

[21] Wagner, H. M., "The Lower Bounded and Partial Upper Bounded Distribution Model," Naval Research Logistics Quarterly, *20*, 265-268 (1973).

A GENERALIZED-INDICES TRANSPORTATION PROBLEM

Deepak Bammi

Champion International
Hamilton, Ohio

ABSTRACT

A generalized-indices transportation problem is formulated and an algorithm is presented for its solution. The algorithm is an extension of the modi-method. A theorem on the number of independent variables in the generalized-indices transportation problem is proved. An example problem is solved for the four-indices transportation problem. A computer program has been written to solve any four-indices problem.

INTRODUCTION

The traditional transportation problem is one of minimizing the distribution costs between a set of origins and a set of destinations. This involves two indices on the variables and the costs. Schell [10] was the first to introduce the concept of a multidimensional transportation problem. In his example there are various types of products to be distributed between a set of origins and a set of destinations. Haley [5] presented a solution procedure for the three-indices problem. In [6] he showed various applications of the three-indices problem. The third index can represent different modes of transport between a set of origins and destinations, or it may represent a set of intermediate warehouses.

Theorems for the existence of a solution to the three-indices problem have been presented by Haley [6, 7, 8], by Moravek and Vlach [9], and by Smith [12]. Corban [2] has also solved the multi-index transportation problem.

An example of the four-indices problem would be the minimization of total transportation costs between m factories, n warehouses, p wholesale outlets, and q retail stores. A typical application of this problem would be in the oil industry, where oil has to be transported from refineries to intermediate storage areas, to district headquarters, and finally to individual service stations.

A GENERALIZED-INDICES TRANSPORTATION PROBLEM

Formulation of Problem

We may write this generalized-indices transportation problem as

$$\text{(1)} \qquad \text{Minimize} \sum_{i=1}^{i=m} \sum_{j=1}^{j=n} \sum_{k=1}^{k=p} \cdots \sum_{l=1}^{l=q} x_{ijk \ldots l}\, c_{ijk \ldots l}$$

(2) subject to
$$\sum_{i=1}^{i=m} x_{ijk\ldots l} = x_{,jk\ldots l}, \qquad \forall j,\ k,\ \ldots,\ l,$$
$$\sum_{j=1}^{j=n} x_{ijk\ldots l} = x_{i,k\ldots l}, \qquad \forall i,\ k,\ \ldots,\ l,$$
$$\sum_{k=1}^{k=p} x_{ijk\ldots l} = x_{ij,\ldots l}, \qquad \forall i,\ j,\ \ldots,\ l,$$
$$\vdots$$
$$\sum_{l=1}^{l=q} x_{ijk\ldots l} = x_{ijk\ldots}, \qquad \forall i,\ j,\ k,\ \ldots,$$

where a comma (,) in the subscript of x indicates a summation over that particular subscript, V is an abbreviation for "over all combinations of". Where

(3)
$$\sum_{j=1}^{j=n} x_{,jk\ldots l} = \sum_{i=1}^{i=m} x_{i,k\ldots l}, \quad V\ k\ \ldots\ l,$$
$$\sum_{k=1}^{k=p} x_{,jk\ldots l} = \sum_{i=1}^{i=m} x_{ij,\ldots l}, \quad V\ j\ \ldots\ l,$$
$$\vdots$$
$$\sum_{l=1}^{l=q} x_{,jk\ldots l} = \sum_{i=1}^{i=m} x_{ijk\ldots} \quad \forall\ jk\ \ldots$$

Constraints (3) represent a set of constraints that relate $x_{,jk\ldots l}$ with $x_{i,k\ldots l}$, $x_{ij,\ldots l}$, ..., and $x_{ijk\ldots}$. Similar constraints exist for $x_{i,k\ldots l}$, $x_{ij,\ldots l}$, ..., and $x_{ijk\ldots}$ with only downward combinations considered so as to eliminate duplicate constraints, and

(4)
$$x_{ijk\ldots l} \geqslant 0 \qquad , \forall i,\ j,\ k,\ \ldots\ l.$$

Solution Procedure

The solution procedure is an extension of the modi-method [13, 5]. The northwest-corner method is used to get an initial solution.

Obtaining an Initial Basic Feasible Solution

It is shown in the appendix that there are $mnp\ \ldots\ q - (m-1)\ (n-1)\ (p-1)\ \ldots\ (q-1)$ independent variables in the generalized indices problem. Using the extended northwest-corner

method an initial solution is obtained with at most $mnp \ldots q - (m-1)\ (n-1)\ (p-1) \ldots (q-1)$ nonzero variables. Such a solution is called a basic solution. Infeasibility because of some variables being negative is treated as in the three-indices problem by Haley [5] and is described below.

Starting in the northwest corner $x_{111\ldots11}$ is set equal to $\min(x_{,11\ldots11}, x_{1,1\ldots11}, \ldots, x_{111\ldots1}, x_{111\ldots1,})$. If the minimum is $x_{,11\ldots11}$ dashes are placed for variables $x_{211\ldots11}, x_{311\ldots11}, \ldots$ to signify a zero level for these variables. Similarly if $x_{1,1\ldots11}$ is the minimum then $x_{121\ldots11}, x_{131\ldots11}, \ldots$ would be set equal to zero, etc.

Next, if $x_{111\ldots21}$ has not been set equal to zero it is equated to $\min(x_{,11\ldots21}, x_{1,1\ldots21}, \ldots, x_{111\ldots,1} - x_{111\ldots11}, x_{111\ldots2,})$. Then $x_{111\ldots31}, x_{111\ldots41}, \ldots$ are allocated till $x_{111\ldots,1}$ is exhausted. Next, if $x_{111\ldots12}$ has not been set equal to zero it is equated to $\min(x_{,11\ldots12}, x_{1,1\ldots12}, \ldots, x_{111\ldots,2}, x_{111\ldots1,} - x_{111\ldots11})$. This is repeated till $x_{111\ldots1,}$ is exhausted. This process is repeated till all the variables have been allocated.

Infeasibility

The procedure for generating an initial solution may lead to some basic variables being negative. To remove this infeasibility an unknown quantity ϕ is added to the basic variable which is at a negative level. Integer multiples ($r_{ijk\ldots l}$) of ϕ are then added to basic or nonbasic variables and subtracted from basic variables according to equation (5),

$$\sum_{i=1}^{i=m} r_{ijk\ldots l} = 0, \quad \sum_{j=1}^{j=n} r_{ijk\ldots l} = 0, \quad \sum_{k=1}^{k=p} r_{ijk\ldots l} = 0, \quad \ldots, \quad \sum_{l=1}^{l=q} r_{ijk\ldots l} = 0. \tag{5}$$

This ensures complicance with the feasibility constraints of equation (2).

We then set ϕ equal to $\min(x_{ijk\ldots l}/r_{ijk\ldots l})$ for those basic variables (occupied cells) which have $r_{ijk\ldots l}\ \phi$ subtracted from them. This ensures that no new basic variable becomes negative and makes the largest change to the variable at the negative level. We then add $r_{ijk\ldots l}\ \phi$ to or subtract it from basic or nonbasic variables to obtain the next tableau.

This process is repeated till all basic variables reach a positive level (each variable at a negative level is treated in turn).

An example infeasible solution for a four-indices problem is shown in Tables 1 and 2. The variable x_{1111} is at a negative level of -1. In order to resolve infeasibility, ϕ is added to x_{1111}. Since ϕ has been added to the first row, we must also subtract the same amount from one of the basic variables in the first row; thus ϕ is subtracted from x_{1211}. This is continued till equation (5) is satisfied for each $i, j, k, \ldots, l$.

Next the value of ϕ is determined to be the minimum of (4/1, 2/1, 6/1, 6/1, 1/1, 1/1, 1/1, 2/1). This sets ϕ at 1. Using this value of ϕ we transform to obtain the tableau for the next iteration. In case of degeneracy the newly feasible variable (in this case x_{1111}) is kept at a zero level and the other basic variables which went to zero (in this case $x_{11112}, x_{1222}, x_{2212}$) are set equal to an infinitesimally small quantity ϵ.

Evaluating Dual Costs

For the N-indices problem each occupied cell has N dual costs associated with it, $u_{jk\ldots l}, v_{ik\ldots l}, w_{ij\ldots l}, \ldots, y_{ijk\ldots}$. The sum of these dual costs equals the cost of the variable. Thus

TABLE 1. *Example 1. Resolution of infeasibility in four-indices problem.*
Iteration 1. Fourth subscript l = 1

	$j=1$	$j=1$	$j=2$	$j=2$
$i=1$	$-1+\phi$ $k=1$		$4-\phi$ $k=1$	
$i=1$		$2-\phi$ $k=2$		$2+\phi$ $k=2$
$i=2$	$6-\phi$ $k=1$		$1+\phi$ $k=1$	
$i=2$		$0+\phi$ $k=2$		$6-\phi$ $k=2$

allocation

TABLE 2. *Example 1. Resolution of infeasibility in four-indices problem.*
Iteration 1. Fourth subscript l = 2

	$j=1$	$j=1$	$j=2$	$j=2$
$i=1$	$1-\phi$ $k=1$		$6+\phi$ $k=1$	
$i=1$		$1+\phi$ $k=2$		$1-\phi$ $k=2$
$i=2$	$4+\phi$ $k=1$		$1-\phi$ $k=1$	
$i=2$		$2-\phi$ $k=2$		$6+\phi$ $k=2$

$$u_{jk\ldots l} + v_{ik\ldots l} + w_{ij\ldots l} + \ldots + y_{ijk\ldots} = c_{ijk\ldots l}. \tag{6}$$

There are a total of $mnp\ldots + mn\ldots q + mp\ldots q + \ldots + np\ldots q$ dual costs ($mnp\ldots$ dual costs of the type $y_{ijk\ldots}$, etc.). The number of occupied cells (basic variables) has been shown to be $mnp\ldots q - (m-1)\,(n-1)\,(p-1)\ldots(q-1)$. Thus, the number of dual costs to be set equal to zero is $mnp\ldots + mn\ldots q + mp\ldots q + \ldots + np\ldots q - [mnp\ldots q - (m-1)\,(n-1)\,(p-1)\ldots(q-1)]$. The set of dual costs to be set equal to zero is chosen such that equation (6) is satisfied for each of the basic variables.

Evaluating Unoccupied Cells

For each unoccupied cell the quantity $q_{ijk\ldots l}$ is calculated as the difference between the cost of the variable and the sum of dual costs:

(7) $$q_{ijk\ldots l} = c_{ijk\ldots l} - (u_{jk\ldots l} + v_{ik\ldots l} + w_{ij\ldots l} + \ldots + y_{ijk\ldots}).$$

If some $q_{ijk\ldots l}$ are negative, then the variable with the most negative $q_{ijk\ldots l}$ is introduced into the basis. If all $q_{ijk\ldots l}$ are nonnegative then an optimal solution has been obtained. If in addition some q_{ijkl} are zero, then multiple optima exist, and these can be obtained by entering the relevant variables into the basis.

Altering a Basic Feasible Solution

An unknown quantity θ is added to the entering variable. Integer multiples ($n_{ijk\ldots l}$) of θ are then added to or subtracted from the occupied cells according to equation (8) to maintain feasibility and to continue to satisfy equation (2):

(8) $$\sum_{i=1}^{i=m} n_{ijk\ldots l} = 0, \quad \sum_{j=1}^{j=n} n_{ijk\ldots l} = 0, \quad \sum_{k=1}^{k=p} n_{ijk\ldots l} = 0, \quad \ldots, \quad \sum_{l=1}^{l=q} n_{ijk\ldots l} = 0.$$

The altering of a basic feasible solution with the use of multiples of θ proceeds in a manner similar to removing infeasibilities according to equation (5).

Obtaining a New Basic Feasible Solution

We set θ equal to $\min(x_{ijk\ldots l}/n_{ijk\ldots l})$ for those occupied cells which have $n_{ijk\ldots l}$ θ subtracted from them. This ensures that all basic variables remain nonnegative [equation (4)] while making the largest possible reduction in the objective function. We then add $n_{ijk\ldots l}$, θ to or subtract it from basic variables to obtain the next tableau.

These steps are repeated till an optimal tableau is obtained with all $q_{ijk\ldots l}$ nonnegative. The validity of this procedure follows from a extension of the theorems proved by Haley (6).

EXAMPLE PROBLEM

As an application of this solution procedure an example problem is solved which involves four indices.

Formulation

We can write this example problem as

(9) $$\text{minimize} \sum_{i=1}^{i=3} \sum_{j=1}^{j=2} \sum_{k=1}^{k=2} \sum_{l=1}^{l=2} x_{ijkl}\, c_{ijkl}$$

subject to

(10)
$$\sum_{i=1}^{i=3} x_{i111} = 13, \quad \sum_{i=1}^{i=3} x_{i121} = 7, \quad \sum_{i=1}^{i=3} x_{i211} = 15, \quad \sum_{i=1}^{i=3} x_{i221} = 15,$$
$$\sum_{i=1}^{i=3} x_{i112} = 14, \quad \sum_{i=1}^{i=3} x_{i122} = 16, \quad \sum_{i=1}^{i=3} x_{i212} = 19, \quad \sum_{i=1}^{i=3} x_{i222} = 18,$$
$$\sum_{j=1}^{j=2} x_{1j11} = 7, \quad \sum_{j=1}^{j=2} x_{1j21} = 3, \quad \sum_{j=1}^{j=2} x_{2j11} = 11, \quad \sum_{j=1}^{j=2} x_{2j21} = 10,$$
$$\sum_{j=1}^{j=2} x_{3j11} = 10, \quad \sum_{j=1}^{j=2} x_{3j21} = 9, \quad \sum_{j=1}^{j=2} x_{1j12} = 9, \quad \sum_{j=1}^{j=2} x_{1j22} = 9,$$

$$\sum_{j=1}^{j=2} x_{2j12} = 15,\ \sum_{j=1}^{j=2} x_{2j22} = 10,\ \sum_{j=1}^{j=2} x_{3j12} = 9,\ \sum_{j=1}^{j=2} x_{3j22} = 15,$$

$$\sum_{k=1}^{k=2} x_{11k1} = 5,\ \sum_{k=1}^{k=2} x_{12k1} = 5,\ \sum_{k=1}^{k=2} x_{21k1} = 9,\ \sum_{k=1}^{k=2} x_{22k1} = 12,$$

$$\sum_{k=1}^{k=2} x_{31k1} = 6,\ \sum_{k=1}^{k=2} x_{32k1} = 13,\ \sum_{k=1}^{k=2} x_{11k2} = 6,\ \sum_{k=1}^{k=2} = x_{12k2}\ 12,$$

$$\sum_{k=1}^{k=2} x_{21k2} = 14,\ \sum_{k=1}^{k=2} x_{22k2} = 11,\ \sum_{k=1}^{k=2} x_{31k2} = 10,\ \sum_{k=1}^{k=2} x_{32k2} = 14,$$

$$\sum_{l=1}^{l=2} x_{111l} = 5,\ \sum_{l=1}^{l=2} x_{121l} = 11,\ \sum_{l=1}^{l=2} x_{112l} = 6,\ \sum_{l=1}^{l=2} x_{122l} = 6,$$

$$\sum_{l=1}^{l=2} x_{211l} = 14,\ \sum_{l=1}^{l=2} x_{221l} = 12,\ \sum_{l=1}^{l=2} x_{212l} = 9,\ \sum_{l=1}^{l=2} x_{222l} = 11,$$

$$\sum_{l=1}^{l=2} x_{311l} = 8,\ \sum_{l=1}^{l=2} x_{321l} = 11,\ \sum_{l=1}^{l=2} x_{312l} = 8,\ \sum_{l=1}^{l=2} x_{322l} = 16,$$

where $x_{,jkl}$, $x_{i,kl}$, $x_{ij,l}$, and $x_{ijk,}$ satisfy the following equalities:

(11)

$$\sum_{j=1}^{j=2} x_{,j11} = \sum_{i=1}^{i=3} x_{i,11},\ \sum_{j=1}^{j=2} x_{,j21} = \sum_{i=1}^{i=3} x_{i,21},$$

$$\sum_{j=1}^{j=2} x_{,j12} = \sum_{i=1}^{i=3} x_{i,12},\ \sum_{j=1}^{j=2} x_{,j22} = \sum_{i=1}^{i=3} x_{i,22},$$

$$\sum_{k=1}^{k=2} x_{,1k1} = \sum_{i=1}^{i=3} x_{i1,1},\ \sum_{k=1}^{k=2} x_{,2k1} = \sum_{i=1}^{i=3} x_{i2,1},$$

$$\sum_{k=1}^{k=2} x_{,1k2} = \sum_{i=1}^{i=3} x_{i1,2},\ \sum_{k=1}^{k=2} x_{,2k2} = \sum_{i=1}^{i=3} x_{i2,2},$$

$$\sum_{l=1}^{l=2} x_{,11l} = \sum_{i=1}^{i=3} x_{i11},\ \sum_{l=1}^{l=2} x_{,12l} = \sum_{i=1}^{i=3} x_{i12},$$

$$\sum_{l=1}^{l=2} x_{,21l} = \sum_{i=1}^{i=3} x_{i21},\ \sum_{l=1}^{l=2} x_{,22l} = \sum_{i=1}^{i=3} x_{i22},$$

$$\sum_{k=1}^{k=2} x_{1,k1} = \sum_{j=1}^{j=2} x_{1j,1},\ \sum_{k=1}^{k=2} x_{2,k1} = \sum_{j=1}^{j=2} x_{2j,1},$$

$$\sum_{k=1}^{k=2} x_{3,k1} = \sum_{j=1}^{j=2} x_{3j,1},\ \sum_{k=1}^{k=2} x_{1,k2} = \sum_{j=1}^{j=2} x_{1j,2},$$

$$\sum_{k=1}^{k=2} x_{2,k2} = \sum_{j=1}^{j=2} x_{2j,2},\ \sum_{k=1}^{k=2} x_{3,k2} = \sum_{j=1}^{j=2} x_{3j,2},$$

$$\sum_{l=1}^{l=2} x_{1,1l} = \sum_{j=1}^{j=2} x_{1j1,},\ \sum_{l=1}^{l=2} x_{1,2l} = \sum_{j=1}^{j=2} x_{1j2,},$$

$$\sum_{l=1}^{l=2} x_{2,1l} = \sum_{j=1}^{j=2} x_{2j1,},\ \sum_{l=1}^{l=2} x_{2,2l} = \sum_{j=1}^{j=2} x_{2j2,},$$

$$\sum_{l=1}^{l=2} x_{3,1l} = \sum_{j=1}^{j=2} x_{3j1,},\ \sum_{l=1}^{l=2} x_{3,2l} = \sum_{j=1}^{j=2} x_{3j2,},$$

$$\sum_{l=1}^{l=2} x_{11,l} = \sum_{k=1}^{k=2} x_{11k,}, \quad \sum_{l=1}^{l=2} x_{12,l} = \sum_{k=1}^{k=2} x_{12k,}$$

$$\sum_{l=1}^{l=2} x_{21,l} = \sum_{k=1}^{k=2} x_{21k,}, \quad \sum_{l=1}^{l=2} x_{22,l} = \sum_{k=1}^{k=2} x_{22k,}$$

$$\sum_{l=1}^{l=2} x_{31,l} = \sum_{k=1}^{k=2} x_{31k,}, \quad \sum_{l=1}^{l=2} x_{32,l} = \sum_{k=1}^{k=2} x_{32k,}.$$

Also,

$$x_{ijkl} \geqslant 0 \qquad Vi,j,k,l.$$

Solution

Following the procedure outlined above, an initial basic feasible solution to the example problem is presented in Tables 3, 4, and 5. The variables are allocated according to equation (10) and the end requirements satisfy equation (11). The cost of allocations in the initial basic feasible solution is given by equation (9) to be 441.

TABLE 3. *Example problem. Initial basic feasible solution. Fourth subscript l = 1. Note m = 3, n = 2, p = 2, q = 2.*

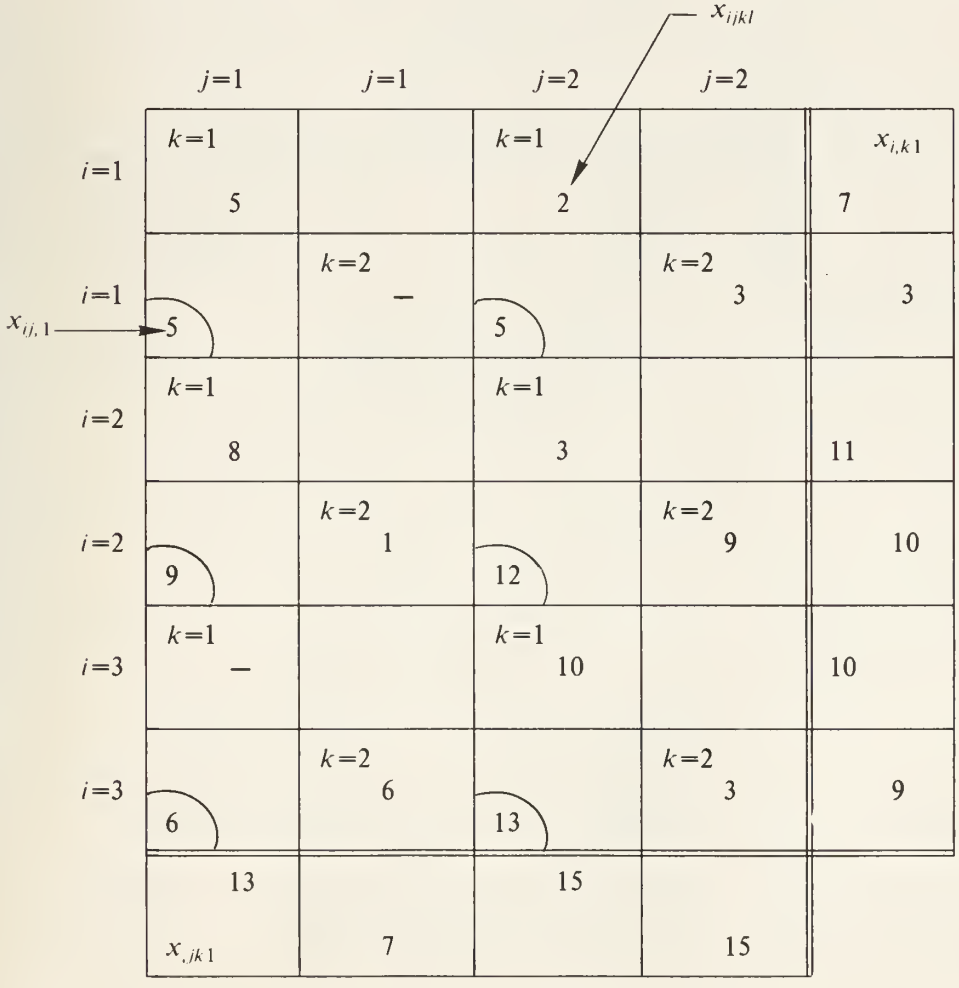

TABLE 4. *Example problem. Initial basic feasible solution. Fourth subscript l = 2.*

	$j=1$	$j=1$	$j=2$	$j=2$	
$i=1$	$k=1$ —		$k=1$ 9		$x_{i,k2}$ 9
$i=1$		$k=2$ 6		$k=2$ 3	9
$x_{ij,2}$ →	6		12		
$i=2$	$k=1$ 6		$k=1$ 9		15
$i=2$		$k=2$ 8		$k=2$ 2	10
	14		11		
$i=3$	$k=1$ 8		$k=1$ 1		9
$i=3$		$k=2$ 2		$k=2$ 13	15
	10		14		
	14		19		
	$x_{,jk2}$	16		18	

TABLE 5. *Example problem. Initial basic feasible solution. Summation over all values of fourth subscript l = 1,2*

i	j	k	x_{ijkl}	x_{ijk2}	$x_{ijk.}$
1	1	1	5	—	5
1	2	1	2	9	11
1	1	2	—	6	6
1	2	2	3	3	6
2	1	1	8	6	14
2	2	1	3	9	12
2	1	2	1	8	9
2	2	2	9	2	11
3	1	1	—	8	8
3	2	1	10	1	11
3	1	2	6	2	8
3	2	2	3	13	16

For this problem $m = 3$, $n = 2$, $p = 2$, and $q = 2$. Thus there are ($mnpq$=) 24 variables and [$mnpq - (m-1)\ (n-1)\ (p-1)\ (q-1)$=] 22 of these are in the basis at a given time. There are a total of ($mnp + mnq + mpq + npq$ =) 44 dual costs and since there are a total of 22 independent variables, (44-22=) 22 of the dual costs may be set equal to zero. The set of dual costs to be set equal to zero is chosen such that equation (6) is satisfied for each of the basic variables.

Work sheets for proceeding from one iteration to the next are shown in Tables 6 and 7. The work sheets display the values of the basic variables x_{ijkl}, cost of each variable c_{ijkl}, dual costs u_{jkl}, v_{ikl}, w_{ijl}, and y_{ijk}, as well as q_{ijkl} (the difference between cost of the nonbasic variable and the sum of dual costs). Of the two nonbasic variables x_{3111} and x_{1112} only the latter has q_{ijkl} negative. Hence x_{1112} enters the basis in the next iteration.

An unknown quantity θ is added to x_{1112} and $n_{ijkl}\,\theta$ is added to and subtracted from some of the occupied cells according to equation (8). Then the value of θ is determined from those cells which have a multiple of θ subtracted from them:

$$\theta = \min(x_{ijkl}/n_{ijkl})$$
$$= \min\ (5/1,\ 3/1,\ 3/1,\ 1/1,\ 9/1,\ 6/1,\ 6/1,\ 2/1)$$
$$= 1$$

TABLE 6. *Example Problem.*
Iteration 1.
Fourth subscript $l = 1$

	$j = 1$ $u_{111} = 0$	$j = 1$ $u_{121} = 0$	$j = 2$ $u_{211} = 0$	$j = 2$ $u_{221} = 0$
$i = 1$ $v_{111} = 0$	$x_{1111} = 5 - \theta$ $w_{111} = 3$ $y_{111} = 3$ $c_{1111} = 6$		$x_{1211} = 2 + \theta$ $w_{121} = 2$ $y_{121} = 4$ $c_{1211} = 6$	
$i = 1$ $v_{121} = 0$		$x_{1121} = \epsilon + \theta$ $w_{111} = 3$ $y_{112} = 5$ $c_{1121} = 8$		$x_{1221} = 3 - \theta$ $w_{121} = 2$ $y_{122} = 3$ $c_{1221} = 5$
$i = 2$ $v_{211} = 0$	$x_{2111} = 8 + \theta$ $w_{211} = 1$ $y_{211} = 3$ $c_{2111} = 4$		$x_{2211} = 3 - \theta$ $w_{221} = 10$ $y_{221} = -7$ $c_{2211} = 3$	
$i = 2$ $v_{221} = 0$		$x_{2121} = 1 - \theta$ $w_{211} = 1$ $y_{212} = 1$ $c_{2121} = 2$		$x_{2221} = 9 + \theta$ $w_{221} = 10$ $y_{222} = -6$ $c_{2221} = 4$
$i = 3$ $v_{311} = -8$	$q_{3111} = 10$ $w_{311} = 0$ $y_{311} = 7$ $c_{3111} = 9$		$x_{3211} = 10$ $w_{321} = 1$ $y_{322} = 8$ $c_{3211} = 1$	
$i = 3$ $v_{321} = 0$		$x_{3121} = 6$ $w_{311} = 0$ $y_{312} = 6$ $c_{3121} = 6$		$x_{3221} = 3$ $w_{321} = 1$ $y_{322} = 1$ $c_{3221} = 2$

TABLE 7. *Example Problem.*
Iteration 1.
Fourth subscript l = 2

	$j = 1$, $u_{112} = 0$	$j = 1$, $u_{122} = 0$	$j = 2$, $u_{212} = 0$	$j = 2$, $u_{222} = 0$
$i = 1$, $v_{112} = 0$	$q_{1112} = -1$ $w_{112} = 0$ $y_{111} = 3$ $c_{1112} = 2$		$x_{1212} = 9 - \theta$ $w_{122} = 0$ $y_{121} = 4$ $c_{1212} = 4$	
$i = 1$, $v_{122} = 0$		$x_{1122} = 6 - \theta$ $w_{112} = 0$ $y_{112} = 5$ $c_{1122} = 5$		$x_{1222} = 3 + \theta$ $w_{122} = 0$ $y_{122} = 3$ $c_{1222} = 3$
$i = 2$, $v_{212} = -8$	$x_{2112} = 6 - \theta$ $w_{212} = 7$ $y_{211} = 3$ $c_{2112} = 2$		$x_{2212} = 9 + \theta$ $w_{222} = 16$ $y_{221} = -7$ $c_{2212} = 1$	
$i = 2$, $v_{222} = 0$		$x_{2122} = 8 + \theta$ $w_{212} = 7$ $y_{212} = 1$ $c_{2122} = 8$		$x_{2222} = 2 - \theta$ $w_{222} = 16$ $y_{222} = -6$ $c_{2222} = 10$
$i = 3$, $v_{312} = 0$	$x_{3112} = 8$ $w_{312} = -3$ $y_{311} = 7$ $c_{3112} = 4$		$x_{3212} = 1$ $w_{322} = 1$ $y_{321} = 8$ $c_{3212} = 9$	
$i = 3$, $v_{322} = 0$		$x_{3122} = 2$ $w_{312} = -3$ $y_{312} = 6$ $c_{3122} = 3$		$x_{3222} = 13$ $w_{322} = 1$ $y_{322} = 1$ $c_{3222} = 2$

Next $n_{ijkl}\theta$ is added to or subtracted from basic variables to yield Tables 8 and 9 of iteration 2. The dual costs are evaluated as before. Since the q_{ijkl} associated with both the nonbasic variables x_{2121} and x_{3111} are nonnegative, an optimal solution has been obtained. Since both the q_{ijkl} are positive, we have a unique optimal point. The minimum cost is 440.

A computer program has been written to solve any four-indices transportation problem.

Interpretation of End Requirements

For a four-indices transportation problem involving minimization of transportation costs between factory i, warehouse j, wholesale outlet k, and retail store l, the following definitions are valid for the end requirements $x_{,jkl}$, $x_{i,kl}$, $x_{ij,l}$, and $x_{ijk,}$:

$x_{,jkl}$ = amount shipped from warehouse j via wholesaler k to retailer l,
$x_{i,kl}$ = amount shipped from factory i via wholesaler k to retailer l,
$x_{ij,l}$ = amount shipped from factory i via warehouse j to retailer l,
$x_{ijk,}$ = amount shipped from factory i via warehouse j to wholesaler k.

These end requirements can be determined for any practical problem from the total supply at factory i, inventory capacity of warehouse j, inventory capacity of wholesaler k, total requirement at retailer l, and equation (3).

TABLE 8. *Example Problem.*
Iteration 2.
Fourth subscript $l = 2$

	$j = 1$ $u_{111} = 0$	$j = 1$ $u_{121} = 0$	$j = 2$ $u_{211} = 0$	$j = 2$ $u_{221} = 1$
$i = 1$ $v_{111} = 0$	$x_{1111} = 4$ $w_{111} = 0$ $y_{111} = 6$ $c_{1111} = 6$		$x_{1211} = 3$ $w_{121} = 0$ $y_{121} = 6$ $c_{1211} = 6$	
$i = 1$ $v_{121} = 1$		$x_{1121} = 1$ $w_{111} = 0$ $y_{112} = 7$ $c_{1121} = 8$		$x_{1221} = 2$ $w_{121} = 0$ $y_{122} = 3$ $c_{1221} = 5$
$i = 2$ $v_{211} = 1$	$x_{2111} = 9$ $w_{211} = 2$ $y_{211} = 1$ $c_{2111} = 4$		$x_{2211} = 2$ $w_{221} = 1$ $y_{221} = 1$ $c_{2211} = 3$	
$i = 2$ $v_{221} = 1$		$q_{2121} = 1$ $w_{211} = 2$ $y_{212} = -2$ $c_{2121} = 2$		$x_{2221} = 10$ $w_{221} = 1$ $y_{222} = 1$ $c_{2221} = 4$
$i = 3$ $v_{311} = 1$	$q_{3111} = 9$ $w_{311} = 0$ $y_{311} = -1$ $c_{3111} = 9$		$x_{3211} = 10$ $w_{321} = 0$ $y_{322} = 0$ $c_{3211} = 1$	
$i = 3$ $v_{321} = -1$		$x_{3121} = 6$ $w_{311} = 0$ $y_{312} = 7$ $c_{3121} = 6$		$x_{3221} = 3$ $w_{321} = 0$ $y_{322} = 2$ $c_{3221} = 2$

TABLE 9. *Example Problem.*
Iteration 2.
Fourth subscript l = 2

	$j = 1$ $u_{112} = 0$	$j = 1$ $u_{122} = 0$	$j = 2$ $u_{212} = 0$	$j = 2$ $u_{222} = 0$
$i = 1$ $v_{112} = 0$	$x_{1112} = 1$ $w_{112} = -4$ $y_{111} = 6$ $c_{1112} = 2$		$x_{1212} = 8$ $w_{122} = -2$ $y_{121} = 6$ $c_{1212} = 4$	
$i = 1$ $v_{122} = 2$		$x_{1122} = 5$ $w_{112} = -4$ $y_{112} = 7$ $c_{1122} = 5$		$x_{1222} = 4$ $w_{122} = -2$ $y_{122} = 3$ $c_{1222} = 3$
$i = 2$ $v_{212} = 1$	$x_{2112} = 5$ $w_{212} = 0$ $y_{211} = 1$ $c_{2112} = 2$		$x_{2212} = 10$ $w_{222} = -1$ $y_{221} = 1$ $c_{2212} = 1$	
$i = 2$ $v_{222} = 10$		$x_{2122} = 9$ $w_{212} = 0$ $y_{212} = -2$ $c_{2122} = 8$		$x_{2222} = 1$ $w_{222} = -1$ $y_{222} = 1$ $c_{2222} = 10$
$i = 3$ $v_{312} = 9$	$x_{3112} = 8$ $w_{312} = -4$ $y_{311} = -1$ $c_{3122} = 4$		$x_{3212} = 1$ $w_{322} = 0$ $y_{321} = 0$ $c_{3212} = 9$	
$i = 3$ $v_{322} = 0$		$x_{3122} = 2$ $w_{312} = -4$ $y_{312} = 7$ $c_{3122} = 3$		$x_{3222} = 13$ $w_{322} = 0$ $y_{322} = 2$ $c_{3222} = 2$

For another four-indices transportation problem involving minimization of transportation costs between factory i and retail store l for product j and by mode of transportation k, the following definitions would apply for $x_{,jkl}$, $x_{i,kl}$, $x_{ij,l}$, and $x_{ijk,}$:

$x_{,jkl}$ = amount shipped of product j by mode k to retailer l,
$x_{i,kl}$ = amount shipped from factory i by mode k to retailer l,
$x_{ij,l}$ = amount shipped from factory i by product j to retailer l,
$x_{ijk,}$ = amount shipped from factory i of product j by mode k.

APPENDIX

For the *two*-indices transportation problem Hadley [4], Dantzig [3], and Simonnard and Hadley [11] have all shown that the number of independent variables for a problem with m sources and n destinations is $(m + n - 1)$. This is equivalent to $\{1 + (m-1) + (n-1)\}$ independent variables.

Schell [10] and Bammi [1] have shown that the number of independent variables in a *three*-indices transportation problem with m factories, n warehouses, and p types of products or p retail stores is $\{mnp - (m-1)\ (n-1)\ (p-1)\}$. This is equivalent to $\{1 + (m-1) + (n-1) + (p-1) + (m-1)\ (n-1) + (m-1)\ (p-1) + (n-1)\ (p-1)\}$ independent variables.

The number of independent variables in a four-indices problem with m factories, n warehouses, p wholesale outlets, and q retail stores is given by $\{mnpq - (m-1)\ (n-1)\ (p-1)\ (q-1)\}$. Changing to a general notation and replacing m by m_1, n by m_2, p by m_3, and q by m_4 it can be stated that the number of independent variables in the four-indices problem equals one plus $\binom{4}{1}$ variables of type $(m_i - 1)$, plus $\binom{4}{2}$ variables of type $(m_i - 1)\ (m_j - 1)$, $j > i$, plus $\binom{4}{3}$ variables of type $(m_i - 1)\ (m_j - 1)\ (m_k - 1)\ (m_k - 1)$, $j > i$, $k > j$.

THEOREM: *The number of independent variables in the N-indices transportation problem is one plus the sum of the products over all combinations of* $(m_i - 1)$ *taken 1, 2, . . . , (N-1) at a time from N.*

Proof of the theorem is by induction. The $(N + 1)$-indices theorem is shown to be true if the N-indices theorem holds, then the N-indices theorem is true if the $(N\text{-}1)$-indices theorem holds, and so on. The four-indices theorem is true if the three-indices theorem holds. The three- and two-indices theorems have already been shown to be true. Therefore the four-, five-, . . . , $(N\text{-}1)$-, and N-indices theorems are also true.

PROOF: It will be shown that the $(N+1)$-indices theorem is true if the N-indices theorem holds.

There are $m_1\ m_2\ m_3 \ldots m_N\ m_{N+1}$ variables in the $(N+1)$-indices problem. All these variables are contained in a hypercuboid of $(N+1)$ sides the dimension of whose sides are $m_1, m_2, m_3, \ldots, m_N, m_{N+1}$ (hereafter called the *first* hypercuboid). Consider the next smaller hypercuboid of $N+1$ sides having dimensions $m_1 - 1, m_2 - 1, m_3 - 1, \ldots, m_N - 1, m_{N+1} - 1$ (hereafter called the *second* hypercuboid) and located at one of the vertices of the *first* hypercuboid. Next consider the $N + 1$ hypercuboids of $N + 1$ sides and dimensions $m_1 - 1, m_2 - 1, m_3 - 1, \ldots, m_{i-1} - 1, 1, m_{i+1} - 1, \ldots, m_N - 1, m_{N+1} - 1$ (referred to as the *third* type of hypercuboid). The variables in these $N + 1$ hypercuboids of the *third* type are dependent on the variables in the *second* hypercuboid. To prove this linear dependence, the N-indices theorem is applied to a multidimensional vector in one of $N + 1$ hypercuboids of the *third* type and the corresponding set of $m_i - 1$ vectors in the *second* hypercuboid. The N-indices theorem is applied in this fashion to each of the $m_j - 1$ vectors in each of the $N + 1$ hypercuboids of the *third* type.

The above application of the N-indices theorem indicates that the number of independent variables in the *second* hypercuboid is

$$\sum_{i=1}^{N+1} (m_1 - 1)\ (m_2 - 1)\ (m_3 - 1) \ldots (m_{i-1} - 1)\ 1\ (m_{i+1} - 1) \ldots (m_{N+1} - 1),$$

where there is one such term for each of the $N + 1$ faces of the *second* hypercuboid.

Consider next the $N + 1$ hypercuboids of $N + 1$ sides and dimensions $m_1 - 1, m_2 - 1, m_3 - 1, \ldots, m_{i-2} - 1, 1, 1, m_{i+1} - 1, \ldots, m_N - 1, m_{N+1} - 1$ (referred to as the *fourth* type of hypercuboid). By application of the $(N - 1)$-indices theorem it can be shown that the number of independent variables in all hypercuboids of the *fourth* type is

$$\sum_{i-1=1}^{N} \sum_{i=1}^{N} (m_1 - 1)\ (m_2 - 1)\ (m_3 - 1) \ldots (m_{i-2} - 1) \cdot 1 \cdot 1 \cdot (m_{i+1} - 1) \ldots (m_{N+1} - 1).$$

In this manner the N-, $(N-1)$-, $(N-2)$-, ..., 3-, and 2-indices theorems are applied to get the total number of independent variables in the $(N+1)$-indices problem. Thus, the $(N+1)$-indices theorem stands proved if the N-indices theorem is true.

Since the 2- and 3-indices theorems are true, the higher-order theorems are also true.

ACKNOWLEDGMENTS

This paper is based on a master's thesis guided by Professor Spencer B. Smith. Mr. Steven W. Kressner was responsible for writing the computer program for the algorithm. Valuable changes were suggested by the referee. Computations were performed at the computer center of the University of Illinois at Chicago Circle.

REFERENCES

[1] Bammi, Deepak, "A General Solution to the Multi-Index Transportation Problem," Unpublished M. S. Thesis, Illinois Institute of Technology, (1968).

[2] Corban, Adrian, "A Multidimensional Transportation Problem," Abstract in OR/MS Executive Science Institute, Inc., 703-704 (1965).

[3] Dantzig, George B., "Application of the Simplex Method to a Transportation Problem," in *Activity Analysis of Production and Allocation* (Wiley, New York), pp. 359-373.

[4] Hadley, G., *Linear Programming* (Addison-Wesley, Reading, Massachusetts, 1963).

[5] Haley, K. B., "The Solid Transportation Problem," Operations Research *10*, 448-463 (1962).

[6] Haley, K. B., "The Multi-Index Transportation Problem," Operations Research *11*, 369-379 (1963).

[7] Haley, K. B., "The Existence of a Solution to the Multi-Index Problem," Operational Research Quarterly *16*, 471-474 (1965).

[8] Haley, K. B., "Note on the Letter by Moraveck and Vlach," Operations Research *15*, 545-546 (1967).

[9] Moravek, J., and Vlach, M., "On the Necessary Conditions for the Existence of the Solution of the Multi-Index Transportation Problem," Operations Research *15*, 542-545 (1967).

[10] Schell, Emil D., "Distribution of a Product by Several Properties," in *Proceedings of the Second Symposium in Linear Programming 2* (Directorate of Management Analysis, DCS/Comptroller H. Q. U. S. A. F., Washington, D. C., 1955).

[11] Simonnard, Michel A., and Hadley, G. F., "Maximum Number of Iterations in the Transportation Problem," Naval Research Logistics Quarterly *6*, 125-129 (1959).

[12] Smith, Graham, "Further Necessary Conditions for the Existence of a Solution to the Multi-Index Problem," Operations Research *21*, 380-386 (1973).

[13] Vajda, S., *Readings in Linear Programming* (Pitman, New York, 1958).

A PROBLEM IN NETWORK INTERDICTION

Bruce Golden

College of Business and Management
University of Maryland
College Park, Maryland

ABSTRACT

Suppose we are given a network $G=(V,E)$ with arc distances and a linear cost function for lengthening arcs. In this note, we consider a network-interdiction problem in which the shortest path from source node s to sink node t is to be increased to at least τ units via a least-cost investment strategy. This problem is shown to reduce to a simple minimum-cost-flow problem. Applications and generalizations are discussed, including the multiple-destination case.

Let $G=(V,E)$ be a directed graph with vertex set V, arc set E, arc distances $d(x,y)$, and per-unit costs $c(x,y)$ for increasing the length or travel discomfort of arc (x,y) . That is, the cost of increasing the length of arc (x,y) by $\delta(x,y)$ is given by $c(x,y)\delta(x,y)$. Nodes s and t are distinguished as source and sink nodes, respectively. In a recent paper, Fulkerson and Harding [4] consider the problem of maximizing the length of the shortest path from s to t given a fixed budget b. They demonstrate that the computation required is equivalent to the parametric solution of a minimum-cost-flow problem. In this note, we focus on a closely related problem in which the length of the shortest path from node s to node t is to be increased to at least τ units via a least-cost investment strategy. This problem is shown to reduce to a simple minimum-cost-flow problem.

An important application of this problem, as mentioned by Fulkerson and Harding [4], concerns an opponent's supply of information, manpower, or equipment from a source $s \in V$ to a destination $t \in V$. Increasing the length of the shortest path to at least τ units ensures a predetermined level of difficulty for the opponent. Equivalently, the problem can be viewed as one concerning network reliability. If each arc (x,y) has a probability $p(x,y)$ of being operative, then the interdiction problem becomes one of guaranteeing at minimal cost that the most reliable path from s to t have no more than a prespecified chance of being operative. This equivalence becomes apparent upon taking the logarithm of the product of the arc reliabilities.

The network interdiction problem is formulated as follows:

$$\min \sum_{(x,y)\in E} c(x,y)\ \delta(x,y), \tag{1.1}$$

$$\pi(t) - \pi(s) \geqslant \tau, \tag{1.2}$$

$$\pi(y) - \pi(x) - \delta(x,y) \leqslant d(x,y),\ (x,y)\in E, \tag{1.3}$$

$$\delta(x,y) \geqslant 0. \tag{1.4}$$

The decision variable $\pi(w)$ is a node potential which represents the length of the shortest path from node s to node w when $\pi(s)=0$; the decision variable $\delta(x,y)$ indicates the nonnegative increase in distance on arc (x,y). Equations (1.3) guarantee that the node potentials reflect shortest path lengths. Equation (1.2) constrains the shortest path from s to t to be no shorter than τ units.

If we assign dual variables z and $f(x,y)$ to equations (1.2) and (1.3) respectively, the linear programming dual of (1.1) - (1.4) becomes

$$\max \left[\tau z - \sum_{(x,y)\in E} f(x,y)\, d(x,y)\right], \tag{2.1}$$

$$\sum_{j\in A(x)} f(x,j) - \sum_{m\in B(x)} f(m,x) = \begin{cases} z \text{ for } x = s, \\ 0 \text{ for } x \in V-\{s,t\}, \\ -z \text{ for } x = t, \end{cases} \tag{2.2}$$

$$0 \leqslant f(x,y) \leqslant c\ (x,y), \tag{2.3}$$

where

$$A(x) = \{j \in V \mid (x,j) \in E\}$$

and

$$B(x) = \{m \in V \mid (m,x) \in E\}.$$

Note that for sufficiently large values of τ, where τz dominates the objective function, the above program seeks a maximal flow solution that minimizes the summation in (2.1). The program imposes a premium of τ units on every unit of flow that reaches node t, and a penalty $d(x,y)$ on every unit of flow on arc (x,y). If we view the objective function (2.1) as an equivalent minimization model, we have a minimum-cost-flow problem where z can be thought of as the flow on the "return arc" from t to s.

The relationship between our network-interdiction problem and the maximum minimum-path problem (*MMPP*) treated by Fulkerson and Harding is rather straightforward. An optimal solution (π^*, δ^*) to problem (1.1) - (1.4) is a feasible solution to the *MMPP* if

$$\sum_{(x,y)\in E} c(x,y)\, \delta^*(x,y) \leqslant b.$$

In this case, when τ is increased until the minimum cost solution is equal to b, we have solved the *MMPP*. If (π^*, δ^*) is not budget feasible, then decreasing τ will solve the *MMPP*.

Now suppose the interdiction task requires that shortest paths from s to nodes $t(1)$, $t(2)$, ..., $t(k)$ be no less in length than $\tau(1)$, $\tau(2)$, ..., $\tau(k)$ units respectively. This multiple destination problem can be written

$$\min \sum_{(x,y)\in E} c(x,y)\, \delta(x,y), \tag{3.1}$$

$$\pi[t(i)] - \pi(s) \geqslant \tau(i), i = 1, 2, \ldots, k, \tag{3.2}$$

$$\pi(y) - \pi(x) - \delta(x,y) \leqslant d(x,y), (x,y) \in E, \tag{3.3}$$

$$\delta(x,y) \geqslant 0. \tag{3.4}$$

The dual problem, given below, is again a minimum-cost-flow problem.

$$\max \left[\sum_{i=1}^{k} \tau(i)\, z(i) - \sum_{(x,y) \in E} f(x,y)\, d(x,y) \right], \tag{4.1}$$

$$\sum_{j \in A(x)} f(x,j) - \sum_{m \in B(x)} f(m,x) = \begin{cases} \sum_{i=1}^{k} z(i) & \text{for } x = s, \\ 0 & \text{for } x \in V - \{s, t(1), \ldots, t(k)\}, \\ -z(i) & \text{for } x = t(i), \end{cases} \tag{4.2}$$

$$0 \leqslant f(x,y) \leqslant c(x,y), \tag{4.3}$$

with $A(x)$ and $B(x)$ defined as before.

The minimum-cost network flow problem has been the focus of much research attention in recent years. For example, Barr et al. [2], Hatch [5], and Aashtiani and Magnanti [1] use special data structures to reduce the computation time of the out-of-kilter primal-dual algorithm by an order of magnitude. Problems with thousands of nodes and arcs are now solved in a matter of seconds. Researchers at the Center for Cybernetic Studies, University of Texas, have also constructed very efficient special-purpose primal simplex codes for solving these problems. Since the network interdiction problem can be viewed as a minimum-cost-flow problem, it can be solved easily and quickly.

REFERENCES

[1] Aashtiani, H., and T. Magnanti, "Implementing Primal-Dual Network Flow Algorithms," Working Paper OR 055-76, Operations Research Center, Massachusetts Institute of Technology, Cambridge, Mass. (revised August 1977).

[2] Barr, R., F. Glover, and D. Klingman, "An Improved Version of the Out-of-Kilter Method and a Comparative Study of Computer Codes," Mathematical Programming *7* (1), 60-87 (1974).

[3] Ford, L.R., Jr. and D.R. Fulkerson, *Flows in Networks*, (Princeton University Press, Princeton, N.J., 1962).

[4] Fulkerson, D.R., and G.C. Harding, "Maximizing the Minimum Source-Sink Path Subject to a Budget Constraint," Mathematical Programming *13* (1), 116-118 (1977).

[5] Hatch, R.S., "Bench Marks Comparing Transportation Codes Based on Primal Simplex and Primal-Dual Algorithms," Operations Research *23* (6), 1167-1172 (1975).

REFINED PREDICTION FOR LINEAR REGRESSION MODELS*

J.L. Hess and R.F. Gunst

Department of Statistics
Southern Methodist University
Dallas, Texas

ABSTRACT

Adequate prediction of a response variable using a multiple linear regression model is shown in this article to be related to the presence of multicollinearities among the predictor variables. If strong multicollinearities are present in the data, this information can be used to determine when prediction is likely to be accurate. A region of prediction, R, is proposed as a guide for prediction purposes. This region is related to a prediction interval when the matrix of predictor variables is of full column rank, but it can also be used when the sample is undersized. The Gorman-Toman ten-variable data is used to illustrate the effectiveness of the region R.

1. INTRODUCTION

Prediction of future observations is one of the primary uses of an estimated linear regression model. Although a large number of papers and books have been written on the analysis of regression data, the emphasis in the literature is heavily weighted toward problems of model building and estimation of model parameters, and not on recommendations for using prediction equations. While these problems are all related, they do not necessarily place the same demands on the estimated model.

Currently much of the statistical literature on linear regression is focussing on properties of biased regression estimators. Notable articles include those of James and Stein [10], Hoerl and Kennard [9], Marquardt [14], Lindley and Smith [12], Hawkins [7], and Webster, Gunst, and Mason [23]. Biased estimation is receiving such prominence due to the realization that multicollinearity among the predictor variables (defined in section 2) tends to severely distort the least-squares estimates of the regression parameters. This in turn can result in poor prediction of future responses. Subset selection procedures likewise are not immune to distortion in the presence of multicollinear data.

Underlying this need for good parameter estimates is the assumption that the fitted model is to be used to predict over a wide region of interest of the predictor variables, perhaps an entire rectangular region defined by the extreme values observed on each predictor variable.

*This research was sponsored in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-75-2871.

This may be an unduly stringent assumption as Hocking [8] discusses from a variable selection viewpoint. In other words, frequently it is not ncessary to predict over such a wide region. When this is so, accurate predictions may be possible despite uncertainties about the goodness of individual parameter estimates.

The purpose of this paper is to better identify when prediction is likely to be accurate with multicollinear data. Specifically, this paper was stimulated by three problems noticed by Owen and Reynolds [18] in their development of a prediction equation for estimating engineering manhours for proposed aircraft programs:

1. they decided to include no more than 12 predictor variables from a total of about 60 possible ones since only 23 observations on the response variable (engineering man-hours) were available;

2. a backward elimination [2, Chapter 6] procedure was performed to reduce further the number of predictor variables to eliminate "the possibility of the accidental deletion of a significant variable due to its interacton with other variables"; and
3. the authors concluded that "some limits of extrapolation for formulas should be a primary objective of future studies."

We will address each of these problems in subsequent sections, not with the goal of providing final, definitive solutions to them, but rather to show how each affects the estimation of the regression parameters and the use of the resulting prediction equation. We do not intend to argue that any particular estimator is the best one to use with multicollinear data. We will, however, point out some advantages of using a principal component estimator to obtain a prediction equation.

2. LEAST-SQUARES PREDICTION EQUATIONS

In this section we will examine problems 2 and 3 of Owen and Reynolds, which were listed in the previous section. Suppose the assumed linear regression model is written as

$$\mathbf{Y} = \beta_0 \vec{1} + X\vec{\beta} + \vec{\epsilon}, \tag{1}$$

where $\mathbf{Y}$ is an $(n \times 1)$ vector of observations on the response (dependent) variable, $\vec{1}$ is an $(n \times 1)$ vector of ones, $X = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]$ is an $(n \times p)$ full column rank matrix of predictor (independent) variables, β_0 is an unknown constant, $\vec{\beta}$ is a $(p \times 1)$ vector of unknown regression parameters, and $\vec{\epsilon}$ is an $(n \times 1)$ vector of unobservable random error terms with $\vec{\epsilon} \sim N(\vec{0}, \sigma^2 I)$. For simplicity, we assume that the elements of $\mathbf{X}_j$ are standardized so that $\mathbf{X}_j' \vec{1} = 0$ and $\mathbf{X}_j' \mathbf{X}_j = 1$ $j = 1, 2, \ldots, P$. Finally, model (1) is assumed to represent adequately the response variable although some of the predictor variables may not be needed for adequate prediction.

The three problems cited by Owen and Reynolds result from inadequacies in the data used to estimate the model parameters, in particular, from multicollinearities in the data. A multicollinearity can be defined as a linear combination of the columns of X that is nearly zero. This implies that $X'X$ is nearly singular. A multicollinearity is not necessarily due to some variables being redundant in the specification of the model, but they may be redundant *for the data collected.*

Redundant model variables, those variables that will be redundant for all samples of data, can and should be deleted from the model since they serve only to inflate the variance of predicted responses (see, e.g., [8]). If the redundancy is inherent only in the particular data sampled, it is dangerous to remove them from the predictor since the estimated model may then be biased when future responses are predicted. Yet multicollinearities tend to cause the deletion of one or more of the multicollinear variables merely because they are involved in multicollinearities, not because they are worthless predictor variables.

To see this latter point, denote the eigenvalues of $X'X$ by $l_1 \geqslant l_2 \geqslant \ldots \geqslant l_p$ and the corresponding eigenvectors by $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_p$. If there are one or more multicollinearities among the columns of X, one or more of the eigenvalues of $X'X$ will be nearly zero. For eigenvalues that are near zero, multicollinearities can be identified by noting that

$$\mathbf{V}_j' X'X\mathbf{V}_j = l_j \approx 0 \rightarrow \sum_{i=1}^{p} V_{ij}\mathbf{X}_j \approx \vec{0}. \tag{2}$$

Equation (2) shows that the eigenvector $\mathbf{V}_j$ corresponding to a small eigenvalue l_j provides the coefficients for the linear combination of the columns of X causing a multicollinearity. Naturally, the larger elements in $\mathbf{V}_j$ identify the predictor variables most strongly multicollinear. Reference [16] contains a more complete discussion of multicollinearities and the problems associated with them.

The least-squares estimator of $\vec{\beta}$ for the model specified by (1) is

$$\hat{\vec{\beta}} = (X'X)^{-1}X'\mathbf{Y}.$$

The variances and covariances of the $\hat{\beta}_i$ can be found from

$$\mathrm{Var}[\hat{\vec{\beta}}] = (X'X)^{-1}\sigma^2 = \sum_{j=1}^{p} l_j^{-1}\mathbf{V}_j\mathbf{V}_j'\sigma^2. \tag{3}$$

From (3) we can see that small eigenvalues in $X'X$ will result in large variances and covariances for estimated parameters of variables involved in multicollinearities (those with large V_{ij} values in (2)).

When attempting to reduce the number of variables in the prediction equation, the t statistic commonly used to test H_0: $\beta_j = 0$ is

$$t = \hat{\beta}_j/(c_{jj}\,\mathrm{MSE})^{1/2}, \tag{4}$$

where c_{jj} is the j th diagonal element of $(X'X)^{-1}$ and MSE is the estimate of σ^2 computed from the full model (1). Since the c_{jj} values of variables involved in multicollinearities tend to be large due to the small l_j in (3), the t statistics corresponding to these variables tend to be small. This accounts for the tendency for variables to be deleted by some computer programs because of their "interaction with other variables".

Contrary to Owen and Reynolds' supposition, backward elimination also suffers from this problem. Backward elimination deletes the variable with the smallest t statistic at each stage. Since multicollinear variables tend to have small t statistics, at least one multicollinear predictor variable is likely to be deleted from the model. See [6] for an illustration of this property.

The problem of eliminating important predictor variables, problem 2 of the previous section, is thus directly related to multicollinearities in the data. Multicollinearities in the data used to estimate $\vec{\beta}$ may affect prediction even if all the predictor variables are used in the prediction equation. Write the least squares prediction equation as

$$\hat{Y} = \hat{\beta}_0 + \mathbf{u}'\vec{\hat{\beta}} \tag{5}$$

where $\beta_0 = \bar{Y}$ and $\mathbf{u}$ is a vector of values of the p-predictor values which are standardized as in (1).

Since $\bar{Y}$ generally estimates β_0 well, (5) will be an adequate predictor of the response if

$$\mathbf{u}'\vec{\hat{\beta}} \approx \mathbf{u}'\vec{\beta}$$

for all values of $\mathbf{u}$ in some region of interest. Now $\mathbf{u}'\vec{\hat{\beta}}$ is an unbiased estimator of $\mathbf{u}'\vec{\beta}$, with variance

$$\operatorname{Var}[\mathbf{u}'\vec{\hat{\beta}}] = \sigma^2\,\mathbf{u}'(X'X)^{-1}\mathbf{u}$$

$$= \sigma^2 \sum_{j=1}^{p} l_j^{-1}\,(\mathbf{u}'\mathbf{V}_j)^2. \tag{6}$$

It can be seen from (6) that $\operatorname{Var}[\mathbf{u}'\vec{\hat{\beta}}]$ will be unacceptably large for many points $\mathbf{u}$ if one or more of the l_j are sufficiently small, or some of the $\mathbf{u}'\mathbf{V}_j$ are large.

A commonly known but infrequently used means of estimating the precision of prediction is to form a 100 $(1-\alpha)$ % prediction interval for Y at the point $\mathbf{u}$:

$$\hat{Y} - t_\nu(\alpha/2)\cdot s \leqslant Y \leqslant \hat{Y} + t_\nu(\alpha/2)\cdot s, \tag{7}$$

where $t_\nu(\alpha/2)$ is the upper 100 $(\alpha/2)$ % point of the t distribution with $\nu = n-p-1$ degrees of freedom, and $s = [(1+n^{-1}+\mathbf{u}'(X'X)^{-1}\mathbf{u}) \cdot \mathrm{MSE}]^{1/2}$. The width of this prediction interval depends on $\sum_{j=1}^{p} l_j^{-1}(\mathbf{u}'\mathbf{V}_j)^2$, as in (6).

Both (6) and (7) essentially depend on how small $\mathbf{u}'\mathbf{V}_j$ is relative to l_j. If l_j is extremely small, the $\mathbf{u}'\mathbf{V}_j$ must also be small or prediction will be poor. These considerations suggest the definition of a "region of predictability" wherein prediction would be expected to be suitably accurate. One such region can be defined as

$$R = \{\mathbf{u}: |\mathbf{u}'\mathbf{V}_j| \leqslant c_j, \quad j=1, \ldots, p, \text{ and } a_i \leqslant u_i \leqslant b_i\}, \tag{8}$$

Where a_i and b_i are the minimum and maximum standardized values of the ith predictor variable, i.e., $a_i = \min\{X_{ki};\ k=1,\ 1,\ \ldots,n\}$ and $b_i = \max\{X_{ki};\ k=1,2,\ \ldots,\ n\}$ for $i=1,2,\ \ldots,\ p$.

Two methods for choosing the c_j are

(i) $c_j = l_j^{1/2}$, or

(ii) $c_j = \max\{|\mathbf{w}_i'\mathbf{V}_j|,\ i=1,2,\ \ldots,\ n\}$ where $\mathbf{w}_i$ is the ith row of X. Method (i) insures that $l_j^{-1}(\mathbf{u}'\mathbf{V}_j)^2 \leqslant 1$, while (ii) bounds $\mathbf{u}'V_j$ by the largest of the values for the points $\mathbf{w}_i'$ used to estimate β. Each of these methods can be interpreted as requiring that the prediction equation be used only in regions for which data has been collected, i.e., the requirements (i) and (ii) limit extrapolation. If one wishes to predict outside R, the predicted values must be cautiously used, but this does indicate a partial response to problem 3 of Owen and Reynolds.

By far the worst prediction will occur for points which have values of $|\mathbf{u}'\mathbf{V}_j|$ that are large for small values of l_j. Suppose r multicollinearities have been detected by a careful examination of the l_j and $\mathbf{V}_j$ and possibly by other procedures such as investigating the "correlation" matrix $X'X$ or the variance inflation factors [14,15]. The c_j, a_i, and b_i could then be relaxed in (8) for the first $(p-r)$ directions $\mathbf{u}'\mathbf{V}_j$, $j=1,2,\ \ldots,p-r$. In these directions extrapolation could be allowed with the understanding that (7) should still provide acceptable bounds. These ideas will become even more important with the discussion of problem 1 in Section 3.

Note that R is based solely on sample information, information available to the data analyst at the time he wishes to make a prediction. If (8) is not satisfied, predicton may — and sometimes will — be accurate since (5) is an unbiased estimator of $\beta_0 + \mathbf{u}'\vec{\beta}$. Prediction for $\mathbf{u} \in R$ provides the assurance that the prediction equation is suitably precise.

If variable-selection procedures are used to reduce the number of variables in the model, prediction will be adequate provided (2) holds for the points at which prediction is desired. This implies that $\mathbf{u}'\mathbf{V}_j \approx 0$ for these new points. But this restriction is again in the form of a region (8) with the c_j chosen suitably small for $j = p-r+1, p-r+2,\ \ldots,\ p$. Thus if a region of predictability of the form (8) is constructed, least-squares estimation and variable-selection techniques will tend to yield prediction equations which are accurate despite the multicollinearities in the data used to estimate the parameters. Outside this region the predictor cannot be expected to perform well due to large variances of the predictor or bias due to erroneously deleting important predictor variables.

3. A PRINCIPAL-COMPONENT PREDICTOR FOR UNDERSIZED SAMPLES

Owen and Reynold's first problem, having to use only about 12 of 60 possible predictor variables in their initial models, results from fewer observations than predictor variables being available for the analysis. The full rank analysis of (1) using least squares requires that $n > p$, a requirement not satisfied by their data.

There is a wide range of model-building problems that could be addressed at this point concerning specification of model (1), but it is not within the scope of this paper to do so. We merely wish to raise the obvious questions regarding the deletion of many potentially valuable predictor variables subjectively, on the basis of a partial analysis of the response and a subset of the predictor variables, or by using a stepwise procedure such as forward selection (see [13] for some objections to this technique for full rank models). One acceptable means of deleting variables prior to an analysis of the complete (assumed correct) model (1) occurs if there are model redundancies, as was mentioned in Section 2.

Rather than demanding a full rank analysis, generalized inverse estimators offer another option. The generalized inverse solution is generaly presented in a discussion of singular X matrices (as in designed experiments) for which $n > p$ (see, e.g., [19], [20], or [22]. While the existence of this estimator of $\vec{\beta}$ and its estimability characteristics are well-known, its potential use with undrsized samples $(n \leqslant p)$ has not been fully explored. An exception to this statement is in the economic literature of simultaneous-equations systems [3,11,21].

The particular generalized inverse estimator we will examine in this section is referred to in the literature (e.g. [17,14]) as a principal components estimator. If $n > p$ and X has rank $p-r$ (corresponding to $l_{p-r+1} = l_{p-r+2} = \ldots = l_p = 0$), the principal component estimator is defined to be

$$\tilde{\vec{\beta}} = (X'X)^{-}X'\mathbf{Y} = V_L L_L^{-1} V_L' X'\mathbf{Y} \tag{9}$$

where the generalized inverse of $X'X$ is $(X'X)^{-} = V_L L_L^{-1} V_L'$ $V_L = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_{p-r}]$, and $L_L = \text{diag}\,(l_1, l_2, \ldots, l_{p-r})$. It is often demonstrated that (9) is the least-squares estimator of $\vec{\beta}$ subject to the constraints $V_0' \vec{\beta} = \vec{0}$, where $V_0 = [\mathbf{V}_{p-r+1}, \mathbf{V}_{p-r+2}, \ldots, \mathbf{V}_p]$.

With undersized samples (i.e., $n \leqslant p$) there are usually s very small eigenvalues of $X'X$ in addition to the r zero ones. We propose, therefore, a generalization of (9) for undersized samples which is of the same form but with $V_L = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_{p-r}]$. Again, it can be shown that this is the least squares estimator of $\vec{\beta}$ subject to the constraints $V_0' \vec{\beta} = \vec{0}$ and $V_s' \vec{\beta} = \vec{0}$, where $V_s = [\mathbf{V}_{p-s-r+1}, \mathbf{V}_{p-s-r+2}, \ldots, \mathbf{V}_{p-r}]$.

Our rationale for using this estimator of $\vec{\beta}$ stems from a different justification than the parameter constraints given above. This justification stresses the use of the actual information provided by the matrix of predictor variables and, as we shall see, again yields guidelines for the use of the resulting predictor.

Let H be an $(n \times n)$ matrix of eigenvectors of XX', partitioned as $H = [H_L : H_s : H_0]$. H_L is $n \times (p-s-r)$ and contains the eigenvectors of XX' corresponding to the eigenvalues in $L_L = \text{diag}\,(l_1, \ldots, l_{p-s-r})$, H_s is $(n \times s)$ and contains the eigenvectors corresponding to the eigenvalues in $L_s = \text{diag}\,(l_{p-s-r+1}, \ldots, l_{p-r})$, and H_0 contains the eigenvectors corresponding to the $n-p+r$ zero eigenvalues. For undersized samples, r, the number of zero eigenvectors of $X'X$, is generally equal to $p-(n-1)$, so that H_0 contains only one vector. Then we can write [4]

$$\begin{aligned} X &= H_0 \Phi V_0' + H_s L_s^{1/2} V_s' + H_L L_L^{1/2} V_L' \\ &= X_0 + X_s + X_L, \end{aligned} \tag{10}$$

where Φ is an $(n-p+r) \times r$ matrix of zeros and $X_0 = H_0 \Phi V_0'$, etc. Since $X_0 = \Phi$ and $X_s \approx \Phi$ (since L_s contans small eigenvalues), we see from (10) that $X \approx X_L$. This emphasizes the point that the entire space of predictor variables has not been sampled, only a subspace that is primarily spanned by the eigenvectors in V_L. Inserting X_L in place of X in (1) and obtaining the principal component solution to the normal equations yields (9) with V_L defined as above. This argument can also be used to justify the use of a principal component estimator for the full rank model if multicollinearities are present since, then, $X = X_s + X_L \approx X_L$.

The principal component prediction equation for undersized samples,

$$\tilde{Y} = \bar{Y} + \mathbf{u}'\tilde{\vec{\beta}}, \tag{11}$$

is biased. The bias of (11) can be written

$$\begin{aligned} B(\tilde{Y}) &= \mathbf{u}'\vec{\beta} - \mathbf{u}' (X_L' X_L)^{-} X_L' X \vec{\beta} \\ &= \mathbf{u}'\vec{\beta} - \mathbf{u}' V_L V_L' \vec{\beta}, \end{aligned} \tag{12}$$

and the variance of $\mathbf{u}'\vec{\tilde{\beta}}$ is

$$\text{Var}[\mathbf{u}'\vec{\tilde{\beta}}] = \mathbf{u}'(X_L'X_L)^{-}\mathbf{u}\sigma^2$$

$$= \sigma^2 \sum_{j=1}^{p-s-r} l_j^{-1}(\mathbf{u}'\mathbf{V}_j)^2. \tag{13}$$

The variance term (13) does not suffer from having small eigenvalues as does (6), but (12) indicates that the predictor is generally biased. Note that if $V_0'\mathbf{u} = \vec{0}$ and $V_s'\mathbf{u} = \vec{0}$, $\mathbf{u}'\vec{\beta} = \mathbf{u}'V_L V_L'\vec{\beta}$, and (11) indeed turns out to be unbiased. This again reflects the fact that prediction should be accurate if we restrict the region of predictability to points in a general region that was actually sampled.

This discussion suggests a region similar to (8) within which prediction could be proposed, but outside of which prediction should not be recommended. Extrapolation can also be allowed in the space spanned by V_L. An illustration of these recommendations is the subject of the next section.

4. AN ASSESSMENT OF R

In this section an example is presented to illustrate the potential benefits of using a region such as R as a guide in predicting. Again, a prediction interval of the form (7) is preferable to R when X is of full column rank and a sufficient number of observations are available to obtain a good estimate of σ^2. Otherwise, R can still be effectively used, as is now demonstrated.

The example concerns the ten-variable data of Gorman and Toman [5]. A detailed analysis of this data, including a listing of the raw data, is given by Daniel and Wood [1]. Two analyses of this data are to be performed here; a full-rank analysis in which 15 of the n=36 data points are used to obtain a predictor (points 1, 3,5, ... , 27, and 29) and an undersized-sample analysis in which only the first 10 of the 36 data points are used. Each predictor is then used to predict the remaining observations.

With the full-rank analysis, the smallest latent root of the standardized $X'X$ matrix is $l_{10} = 0.0050$, with corresponding latent vector

$$\mathbf{V}'_{10} = \begin{matrix} (0.448, & 0.099, & 0.046, & 0.098, & -0.398 \\ 0.752, & 0.072, & -0.183, & -0.104, & 0.064) \end{matrix}$$

From the discussion of Section 2, both (7) and (8) suggest that prediction should not be attempted unless $\mathbf{u}'\mathbf{V}_{10}$ is small. (For simplicity and ease of discussion, we are only considering one small latent root in this analysis. Since $l_9 = 0.063$, we may wish to consider the magnitude of $\mathbf{u}'\mathbf{V}_9$ as well). Usng least squares, a predictor of the form (5) was constructed.

Figure 1 is a plot of $|Y_i - \hat{Y}_i$ vs. $|\mathbf{u}'\mathbf{V}_{10}|$ for the 36-15 = 21 data points not used to estimate the parameters in the prediction equation. The trend is clear: the magnitude of the residuals increases with the magnitude of $\mathbf{u}'\mathbf{V}_{10}$. While some moderate-sized residuals do occur with large magnitudes of $\mathbf{u}'\mathbf{V}_{10}$. there are no large residuals for small magnitudes of $\mathbf{u}'\mathbf{V}_{10}$.

Also evident from Figure 1 is the need to explore possible bounds on $\mathbf{u}'\mathbf{V}_{10}$. The two suggested in Section 2 turn out to be

(i) $|\mathbf{u}'\mathbf{V}_{10}| < l_{10}^{1/2} = 0.071$

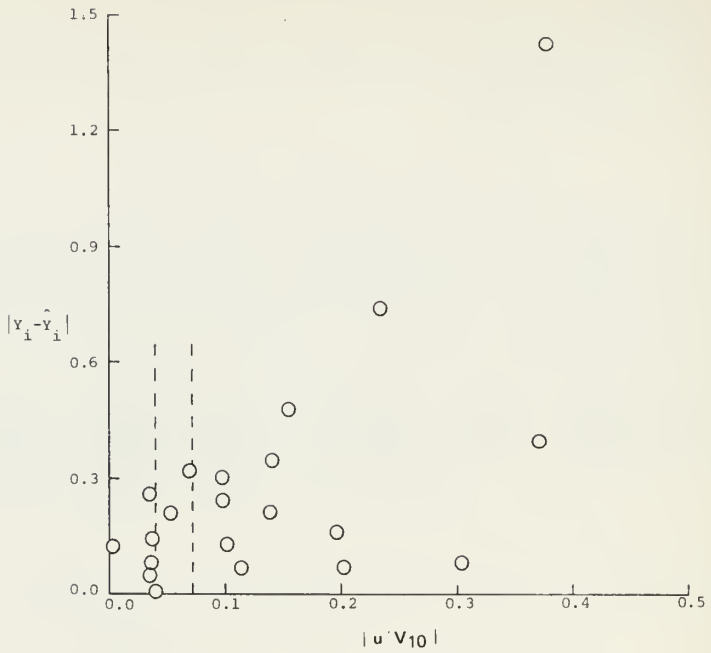

FIGURE 1. Residuals of Gorman-Toman data based on full rank analysis.

and

(ii) $|\mathbf{u}'\mathbf{V}_{10}| < \max\{|\mathbf{w}_i'\mathbf{V}_{10}|\} = 0.039.$

Vertical dashed lines have been drawn in Figure 1 to identify bounds (i) and (ii). Note that (ii) is always more conservative than (i) since

$$l_{10}^{1/2} = \left\{\sum_{i=1}^{n} (w_i'V_{10})^2\right\}^{1/2} \leqslant \max |\mathbf{w}_i'\mathbf{V}_{10}|.$$

For the data plotted in Figure 1, both bounds identify only points for which prediction is relatively good.

For the undersized sample, $l_{10} = 0$ and $l_9 = 0.0104$. The corresponding latent vectors are

$$\mathbf{V}'_9 = \begin{matrix}(0.221, & 0.046, & -0.131, & 0.064, & 0.562\\ -0.645, & -0.092, & 0.148, & 0.219, & -0.342)\end{matrix}$$

and

$$\mathbf{V}'_{10} = \begin{matrix}(0.661, & -0.085, & -0.367, & -0.059, & 0.088\\ 0.495, & 0.290, & 0.188, & 0.160, & -0.139)\end{matrix}$$

The latent vectors corresponding to the eight remaining latent roots were used to estimate $\vec{\beta}$ as in (9) and then form the prediction equation in (11). Figure 2 is a plot of the residuals $|Y_i - \hat{Y}_i|$ of the remaining 36-10 = 26 data points (with circles indicating $|Y_i - \hat{Y}_i| \leqslant 0.75$, triangles indicating $0.75 < |Y_i - \hat{Y}_i| \leqslant 1.50$, and squares indicating $1.50 < |Y_i - \hat{Y}_i|$) as a function of $|\mathbf{u}'\mathbf{V}_{10}|$. Again the trend is clear: smaller residuals occur predominantly with smaller values of both $|\mathbf{u}'\mathbf{V}_9|$ and $|\mathbf{u}'\mathbf{V}_{10}|$.

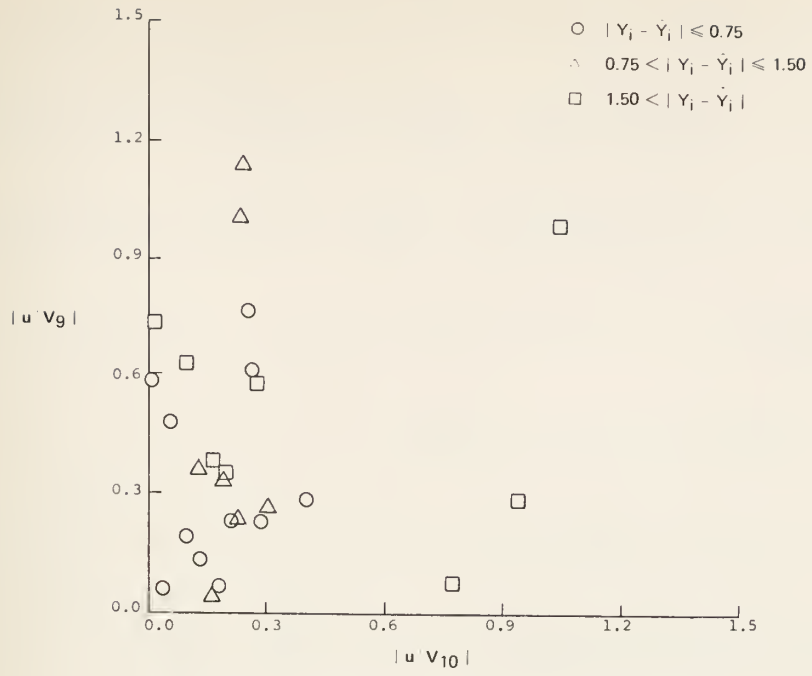

FIGURE 2. Residuals of Gorman-Toman data based on undersized sample analysis.

5. SUMMARY

The intent of this paper is to focus attention on an aspect of regression analysis that is often overlooked when the resulting prediction equation is employed. Regardless of the sample size used to obtain estimates of model parameters (an particularly when the sample size is small), estimation is highly inaccurate outside a region generally defined by (8). Yet regions of this form are always available to the data analyst and can be very valuable as guides in predicting. The Gorman-Toman data illustrates that in both the full-rank situation and the undersized-sample case a region R formed by considering $\mathbf{u}'\mathbf{V}_j$ for latent vectors $\mathbf{V}_j$ corresponding to zero or small latent roots of $X'X$ was effective in identifying when prediction was likely to be inadequate. Further work in this area should concentrate on refining R; in particular, developing reasonable bounds, c_j, for (8) based on the information in X.

REFERENCES

[1] Daniel, C., and F.S. Wood, *Fitting Equations to Data* (Wiley Interscience, New York, 1971).

[2] Draper, N.R., and H. Smith, *Applied Regression Analysis* (Wiley, New York, 1966).

[3] Fisher, W.D., and W.J. Wadycki, "Estimating a Structural Equation in a Large System," Econometrica *39*, 461-65 (1971).

[4] Good, I.J., "Some Applications of the Singular Decomposition of a Matrix," Technometrics *11*, 823-831 (1969).

[5] Gorman, J.W., and R.J. Toman, "Selection of Variables for Fitting Equations to Data", Technometrics *8*, 27-51 (1966).

[6] Gunst, R.F., J.T. Webster, and R.L. Mason, "A Comparison of Least Squares and Latent Root Regression Estimators," Technometrics *18*, 75-83 (1976).

[7] Hawkins, D.M., "On the Investigation of Alternative Regressions by Principal Component Analysis," Applied Statistics *22*, 275-286 (1973).

[8] Hocking, R.R., "The Analysis and Selection of Variables in Linear Regression," Biometrics *32*, 1-49 (1976).

[9] Hoerl, A.E. and R.W. Kennard, "Ridge Regression: Biased Estimation for Non-orthogonal Problems," Technometrics *12*, 55-67 (1970).

[10] James, W., and C. Stein, "Estimation with Quadratic Loss," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability *1*, 361-379 (1961).

[11] Khazzoom, J.D., "An Exploratory Study of the Application of Generalized Inverse to ILS Estimation of Overidentified Equations in Linear Models," Tech. Report No. 219, Department of Statistics, Stanford University, Stanford, California (1975).

[12] Lindley, D.V., and A.F.M. Smith, "Bayes Estimates for the Linear Model," Journal of the Royal Statistical Society, Series B. *34*, 1-18 (1972).

[13] Mantel, N., "Why Stepdown Preocedures in Variable Selection," Technometrics *12*, 621-626 (1970).

[14] Marquardt, D.W., "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," Technometrics *12*, 591-612 (1970).

[15] Marquardt, D.W., and R.D. Snee, "Ridge Regression in Practice," The American Statistician *29*, 3-20 (1975).

[16] Mason, R.L., R.F. Gunst, and J.T. Webster, "Regression Analysis and Problems of Multicollinearity," Communications in Statistics *4*, 277-292 (1975).

[17] Massy, W.F., "Principal Component Regression in Exploratory Statistical Research," Journal of the American Statistical Association *60*, 234-256 (1965).

[18] Owen, D.B., and D.F. Reynolds, "An Application of Statistical Techniques to Estimate Engineering Man-Hours on Major Aircraft Programs," Naval Research Logistics Quarterly *15*, 579-593 (1968).

[19] Rao, C.R., *Linear Statistical Inference and Its Applications* (Wiley, New York, 1965).

[20] Searle, S.R., *Linear Models* (Wiley, New York, 1971).

[21] Swamy, P.A.V.P., and J. Holmes, "The Use of Undersized Samples in the Estimation of Simultaneous Equations Systems," Econometrica *39*, 455-459 (1971).

[22] Theil, H., *Principles of Econometrics* (Wiley, New York, 1971).

[23] Webster, J.T., R.F. Gunst, and R.L. Mason, "Latent Root Regression Analysis," Technometrics *16*, 513-522 (1974).

A NOTE ON SYSTEMS OF SIMULTANEOUS LINEAR DIFFERENCE EQUATIONS

Koji Okuguchi

Department of Economics
Tokyo Metropolitan University
Tokyo

General solutions and specific solution methods for systems of simultaneous linear difference equations with constant coefficients have been provided by Pfouts [3,4]. Consider a reduced system of simultaneous equations with constant coefficients

$$x(t) = B_1 x(t-1) + \ldots + B_m x(t-m), \tag{1}$$

where $x(t) = [x_1(t), \ldots, x_n(t)]'$ and the B_i's are $n \times n$ matrices. Pfouts has shown that (1) can be solved for any variable as

$$x_j(t) = c_1 x_j(t-1) + \ldots + c_r x_j(t-r), \quad r \leqslant mn, \quad j = 1, 2, \ldots, n, \tag{2}$$

where the c_i's do not depend on j. Define an $r \times r$ matrix C derived from the coefficients of the right hand side of (2) as

$$C = \begin{bmatrix} c_1 & c_2 & \cdots & c_r \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ 0 & 0 & \cdots 1 & 0 \end{bmatrix}.$$

Pfouts' Theorem 2 and Theorem 3 together show that $\lim_{t \to \infty} x_j(t) = 0$ for all j if and only if the moduli of characteristic roots of C are all less than unity.

It is clear that we have to calculate the c_i's in order to analyze the stability of (1) by applying Pfouts' method. It is not easy to calculate the c_i's in general, especially when n and/or m are large. This difficulty can be avoided by taking into account the norms of block elements in a partitioned matrix. To see this, first introduce changes of variables defined as

$$Z_1(t) = x(t),$$

$$Z_2(t) = z_1(t-1),$$

$$\ldots\ldots\ldots,$$

$$Z_m(t) = Z_{m-1}(t-1) = x(t-m+1).$$

Then (1) can be rewritten as

$$\begin{bmatrix} Z_1(t) \\ Z_2(t) \\ Z_3(t) \\ \cdot \\ \cdot \\ \cdot \\ Z_m(t) \end{bmatrix} = \begin{bmatrix} B_1 & B_2 \cdots & B_m \\ I & 0 \cdots & 0 \\ 0 & I & 0 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 0 & 0 \cdots I & 0 \end{bmatrix} \begin{bmatrix} Z_1(t-1) \\ Z_2(t-1) \\ Z_3(t-1) \\ \cdot \\ \cdot \\ \cdot \\ Z_m(t-1) \end{bmatrix}. \tag{3}$$

This is the first-order system of simultaneous difference equations with constant coefficients. The coefficient matrix of (3) is the same as B in Pfouts [4]. Let the coefficient matrix of (3) be denoted by B.

Let $A = [A_{IJ}](I, J = 1, 2, \ldots, N)$ be a partitioned matrix whose (I, J) block is A_{IJ}, and let T_A be a matrix whose elements consist of norms of block elements of A, that is, $T_A = [\|A_{IJ}\|]$ $(I, J = 1, 2, \ldots, N)$, where $\|\ \|$ is an arbitrary matrix norm such that $\|I\| = 1$. It has been proven by Okuguchi [2] (see also Conlisk [1]) that the moduli of characteristic roots of A are all less than μ if T_A is indecomposable and $\sum_{J=1}^{N} \|A_{IJ}\| \leqq \mu$, $I = 1, 2, \ldots, N$, with strict inequality for at least one I. Applying this fact to B, the coefficient matrix of (3), we can immediately assert that moduli of characteristic roots of B are all less than unity [hence $\lim_{t\to\infty} Z_i(t) = 0$ for all $i = 1, 2, \ldots, m$, hence $\lim_{t\to\infty} x_j(t) = 0$ by virtue of the definition $Z_1(t) = x(t)$, $j = 1, 2, \ldots, n$] provided that $\|B_i\| \neq 0$, for $i = 1, 2, \ldots, m$, and

$$\sum_{i=1}^{m} \|B_i\| < 1. \tag{4}$$

Clearly the stability condition (4) is easier to apply than Pfouts' which is based on calculation of characteristic roots of C. However, condition (4) is only sufficient for the stability of (1).

REFERENCES

[1] Conlisk, J., "Quick Stability Checks and Matrix Norms," Economica *40*, 402-409 (1973).
[2] Okuguchi, K., "Matrices with Dominant Diagonal Blocks and Economic Theory," Journal of Mathematical Economics *5*, 43-52 (1978).
[3] Pfouts, R. W., and Ferguson, C. E., "A Matrix General Solution of Linear Difference Equations with Constant Coefficients," Mathematics Magazine *33*, 119-127 (1960).
[4] Pfouts, R.W., "A Note on Systems of Simultaneous Linear Differences Equations with Constant Coefficients," Naval Research Logistics Quarterly *12*, 335-340 (1965).

A NOTE ON THE STOCHASTIC SHORTEST-ROUTE PROBLEM

John S. Croucher

Macquarie University
North Ryde, N.S.W.
Australia

ABSTRACT

This paper develops an algorithm for a "shortest route" network problem in which it is desired to find the path which yields the shortest expected distance through the network. It is assumed that if a particular arc is chosen, then there is a finite probability that an adjacent arc will be traversed instead. Backward induction is used and appropriate recursion formulae are developed. A numerical example is provided.

1. INTRODUCTION

Consider a network with nodes numbered 1 through N and an arc $\bar{ij}$ connecting pairs of nodes i and j. Associated with each arc is a nonnegative distance c_{ij}. Assume that the network has no loops; i.e., once a particular node has been left it is impossible to return to it. In a conventional network problem, the objective is to find a shortest path between node 1 and node N. Various solutions to this problem were initially presented by Bellman [1], Dantzig [2], and Dijkstra [3], while Dreyfus [4] has provided a review of some of the more commonly used techniques.

Since that time various extensions and modifications have been made to the problem itself. For example, there may be some cost involved for each distance unit travelled on certain arcs (the objective here being to minimize total cost). More recently, Frieze [5] has examined the case in which path length is defined as a real-valued function defined on the paths. The version considered in this paper is the case in which, once a particular arc is chosen there is a finite probability that an adjacent arc will be traversed instead, i.e., you can't always be certain of travelling along the arc you desire. The aim is to develop an algorithm that will determine the optimal choice of next arc at each node so as to minimize the total expected distance travelled from node 1 to node N.

2. FORMULATION OF THE ALGORITHM

Suppose that at node i ($i = 1, \ldots, N-1$) there are n_i possible arcs which may be traversed. Label their nodes $j_1, j_2, \ldots, j_{n_i}$. Let p_{ij_r} be the probability of traversing the chosen arc $\bar{ij}_r$ ($1 \leqslant r \leqslant n_i$). Assume that if the chosen arc is not travelled, then there is an equal chance that one of the other feasible arcs will be traversed instead. For each of these arcs, this probability is $(1 - p_{ij_r})/(n_i - 1)$ for $n_i > 1$; if $n_i = 1$, then $p_{ij} = 1$.

Let $g(i)$ be the shortest expected distance from node i to node N ($i = 1, \ldots N-1$), and define $g(N) = 0$. Suppose that an arc $\bar{ij}_r$ is *actually* traversed from a node i. Then the shortest expected distance from node i to node N becomes

$$D_{ij_r} = c_{ij_r} + g(j_r).$$

Suppose that it is *decided* to traverse arc $\bar{ij}_r$. Then the shortest expected distance to node N is (for $i = 1, \ldots, N-1$)

$$g(i) = p_{ij_r} D_{ij_r} + \left(\frac{1 - p_{ij_r}}{n_i - 1}\right)(D_{ij_1} + D_{ij_2} + \cdots + D_{ij_{r-1}} + D_{ij_{r+1}} + \cdots + D_{ij_{n_i}}),$$

$$\text{if } n_i > 1,$$

$$= D_{ij_1} \qquad \text{if } n_i = 1.$$

Thus, from node i a succceeding node j is to be found which minimizes $g(i)$ over all possible j. That is for $i = 1, \ldots, N-1$

$$(1) \qquad g(i) = \min_{\substack{jr \\ r=1,\ldots,n_i}} \left\{ p_{ij}[c_{ij_r} + g(j_r)] + \left(\frac{1 - p_{ij_r}}{n_i - 1}\right) \sum_{\substack{k=1 \\ k \neq r}}^{n_i} [c_{ij_k} + g(j_k)] \right\},$$

$$\text{if } n_i > 1,$$

$$= c_{ij_1} + g(j_1) \qquad \text{if } n_i = 1.$$

The equations for $g(i)$ may be solved by backward recursion, i.e., start with the fact that $g(N)$ is zero, then find $g(N-1)$, $g(N-2)$, etc., and finally $g(1)$. An illustration of the technique is given in the next section.

3. NUMERICAL EXAMPLE

Consider a six-node network in which it is desired to travel from node 1 to node 6 covering the shortest distance. The values of the c_{ij} and p_{ij} are given for each arc $\bar{ij}$ in Table 1.

TABLE 1

Arc $\bar{ij}$	c_{ij}	p_{ij}
1-2	2	0.70
1-3	3	0.80
2-3	2	0.40
2-4	3	0.60
2-5	4	0.60
3-4	3	0.60
3-5	3	0.80
4-5	2	0.70
4-6	4	0.60
5-6	1	1.00

The network is displayed diagrammatically in Figure 1, with the arc lengths shown on each arc together with the probability (in parentheses) of traversing that arc.

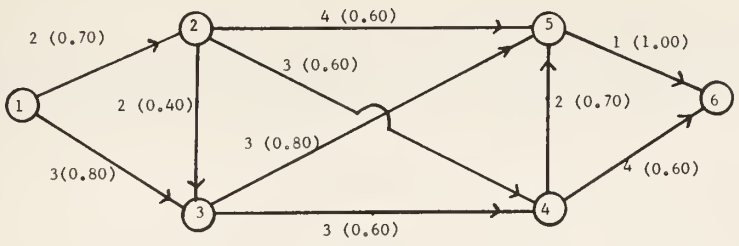

FIGURE 1.

From Table 1 it can be seen that $n_1 = 2$, $n_2 = 3$, $n_3 = 2$, $n_4 = 2$, $n_5 = 1$. The value of $g(i)$ is found for each node, beginning at node 6.

NODE 6: By definition, $g(6) = 0$.

NODE 5: Clearly, $g(5) = 1$, as there is only one possible path to node 6.

NODE 4: Let $j_1 = 5$, $j_2 = 6$. From (1)

$$g(4) = \min_{\substack{j_r \\ r=1,2}} \{p_{4j_r}[c_{4j_r} + g(j_r)] + (1 - p_{4j_r}) \sum_{\substack{k=1 \\ k \neq r}}^{2} [c_{4j_k} + g(j_k)]\}$$

$$= \min(3.30,\ 3.60)$$

$$= 3.30, \textit{ which occurs for } j_1.$$

Thus, at node 4 the next node to aim for is 5, with an expected total distance of 3.30 to node 6.

NODE 3: Let $j_1 = 4$, $j_2 = 5$. From (1),

$$g(3) = \min_{\substack{j_r \\ r=1,2}} \{p_{3j_r}[c_{3j_r} + g(j_r)] + (1 - p_{3j_r}) \sum_{\substack{k=1 \\ k \neq r}}^{2} [c_{3j_k} + g(j_k)]\}$$

$$= \min(5.38,\ 4.46)$$

$$= 4.46, \textit{ which occurs for } j_2.$$

Thus, at node 3 the next node to aim for is 5, with an expected distance of 4.46 to node 6.

NODE 2: Let $j_1 = 3$, $j_2 = 4$, $j_3 = 5$. From (1),

$$g(2) = \min_{\substack{j_r \\ r=1,2,3}} \{p_{2j_r}[c_{2j_r} + g(j_r)] + \frac{(1 - p_{2j_r})}{2} \sum_{\substack{k=1 \\ k \neq r}}^{3} [c_{2j_k} + g(j_k)]\}$$

$$= \min(5.97,\ 6.07,\ 5.55)$$

$$= 5.55, \textit{ which occurs for } j_3.$$

Thus, at node 2 the next node to aim for is node 5, with an expected total distance of 5.55 to node 6.

NODE 1: Let $j_1 = 2$, $j_2 = 3$. From (1),

$$g(1) = \min_{\substack{j_r \\ r=1,2}} \{p_{1j_r}[c_{1j_r} + g(j_r)] + (1 - p_{1j_r}) \sum_{\substack{k=1 \\ k \neq r}}^{2} (c_{1j_k} + g(j_k))\}$$

$$= \min(7.53,\ 7.48)$$

$$= 7.48, \textit{ which occurs for } j_2.$$

Hence, at the source of the network (node 1) the first node to aim for is 3. The total expected distance through the network is 7.48.

4. REMARKS

A dynamic programming approach has been used to solve a stochastic shortest-route problem. The technique used is readily adaptable to solution by computer. It should be noted that the recursion formulae in (1) reduce to those for the conventional shortest-route problem if $p_{ij_r} = 1$ for each pair of nodes i and j_r.

There are a number of applications of the stochastic model, particularly in the field of traffic flow, where it cannot be certain that the chosen path will be the one actually taken. This can apply to everyday situations such as driving (where accidents, road blocks, traffic congestion, etc. may block a desired route) or air travel (where strikes, rerouting, or bad weather may cause a change of path). Other applications are to recreational activities such as bushwalking (bushfires, floods, landslides), hang-gliding (wind variations), and long distance swimming (current changes) and the business activities such as investment over a fixed time horizon where interest rates and investment opportunities are subject to change.

REFERENCES

[1] Bellman, R., "On a Routing Problem," Quarterly of Applied Mathematics *16*, No. 1, 87-90 (1958).

[2] Dantzig, G., "On the Shortest Route through a Network," Management Science *6*, 187-190 (1960).

[3] Dijkstra, E.W., "A Note on Two Problems in Connection with Graphs," Numerische Mathematik *1*, 269-271 (1959).

[4] Dreyfus, S. E., "An Appraisal of Some Shortest-Path Algorithms," Operations Research *17*, 395-412 (1969).

[5] Frieze, A., "Minimum Paths in Directed Graphs," Operational Research Quarterly *28*, No. 2(i), 339-346 (1977).

A NOTE ON THE LIMITATIONS OF GOAL PROGRAMMING AS OBSERVED IN RESOURCE ALLOCATION FOR MARINE ENVIRONMENTAL PROTECTION *

J. Harrald and J. Leotta

Marine Environmental Protection Division
U.S. Coast Guard, Department of Transportation
Washington, D.C.

W.A. Wallace

School of Management
Rensselaer Polytechnic Institute
Troy, New York

R.E. Wendell

Graduate School of Business
University of Pittsburgh
Pittsburgh, Pennsylvania

ABSTRACT

After first formulating the problem of the Marine Environmental Protection program of the Coast Guard as a multiple-objective linear program, we investigate the applicability and limitations of goal programming. We point out how the preemptive goal-programming approach is incompatible with utility preferences. Then we observe the tendency of optimal solutions for standard linear goal programs to occur at extreme points. We also note problems of more general approaches, such as dealing with additively separable approximations to preferences.

INTRODUCTION

In this paper we evaluate the applicability and limitations of goal programming with respect to the Marine Environmental Protection program of the Coast Guard. In the first section we review the Coast Guard problem and formulate it as a multiple-objective mathematical programming problem. Then, we focus on the need for a prior specification of preferences over the activities of the MEP program in order to guide field units and to evaluate the impact of resource constraints. We observe that an attempt to characterize preferences as priorities and then to solve the corresponding preemptive goal-programming problem yields solutions in which priorities are changed and where threshold constraints are introduced. In section 2 we

*This research was supported in part by U.S. Coast Guard, Department of Transportation, under contract No. DOT-CG23876-A.

point out (evidently for the first time) how the preemptive approach is incompatible with preferences representable as a utility function and that, in particular, changes in priorities can be rationalized as a futile attempt to state preferences in a priority fashion. When considering the use of the standard linear goal-programming approach, we note the tendency of optimal solutions to occur at extreme points. We also note problems of more general approaches, such as dealing with additively separable approximations to preferences.

1. THE PROBLEM AND THE ATTEMPTED SOLUTION

The U.S. Federal Water Pollution Control Act (FWPCA) gave the U.S. Coast Guard (USCG) responsibility and enforcement authority for the protection of the marine environment from discharges of oil and other hazardous materials. The Marine Environmental Protection (MEP) program was established by the Coast Guard in 1971 to manage these new duties.

Unfortunately there were no data relating controllable activities such as harbor patrols to any output measures such as the amount of oil spilled. Although a pollution-incident-reporting system was augmented to include data of this type, the problem of determining whether an activity prevents or detects a discharge is still not resolved. For instance, would the number of spills detected increase or decrease with more harbor patrols? Thus, the focus of the Coast Guard became one of identifying those activities that were felt to contribute to the accomplishment of the objectives legislated.

As a result 12 activities, ranging from monitoring the transfer operations of tankships and barges to pursuing a public education program, were selected as surrogates for the "real" objectives of the Coast Guard. Thus, with respect to these surrogates, the Coast Guard viewed the problem as one of having 12 objectives corresponding to these activities. The problem then became one of trying to determine the most effective activity levels subject to resource constraints on personnel, vehicles, and boats [8].

Inasmuch as the most effective activity levels could not be determined by objective methods, the Coast Guard subjectively estimated levels based on the experience of USCG Headquarters and field personnel and on resource-utilization survey data. These subjectively estimated levels were promulgated as Mission Performance Standards in January 1973. Two of them, Standard 1 and Standard 4, are:

S1. Monitoring 25% of all liquid bulk transfer operations involving oil and hazardous substances, ... ;

S4. Conduct daily patrols of the essential harbor area by water once during daylight hours.

These standards became targets for field units in performing their daily activities. Unfortunately, the field units could not feasibly achieve the standards in toto and, in addition, had no guidance in making trade-offs among performance levels of the activities. As a result, it was impossible for MEP program management to estimate how activity levels would change as a function of resources supplied. This posed a difficult situation when the Department of Transportation and the Office of Management and Budget reviewers, who are faced by tight budgetary constraints, tried to determine the marginal utility of any additional resources. This paper discusses, in particular, how goal programming was used to attempt to solve these problems.

However, before we discuss the goal programming approach we give a multiple objective linear programming formulation of a field unit's problem:

(1) $\text{Max}\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\}$

subject to

$$\sum_{j=1}^{12} a_{ij} x_j \leqslant b_i \quad \text{for } i = 1, 2, 3,$$

$$0 \leqslant x_j \leqslant \mu_j \quad \text{for } j = 1, \ldots, 12,$$

where

$x_1 \equiv$ the number of liquid bulk transfer operations involving oil and hazardous substances monitored per period;

$x_2 \equiv$ the number per period of tankships boraded;

$x_3 \equiv$ the number per period of barges boarded;

$x_4 \equiv$ the number per period of daily daylight water patrols of essential harbor areas;

$x_5 \equiv$ the number per period of daily night water patrols of essential harbor areas;

$x_6 \equiv$ the number per period of waterfront facilities spotchecked each month;

$x_7 \equiv$ the number per period of waterfront facilities inspected every six months;

$x_8 \equiv$ the number per period of waterfront facilities to be surveyed once every two years;

$x_9 \equiv$ the number per period of monitors sent to the scene of oil and hazardous substance discharges to determine adequate removal;

$x_{10} \equiv$ the number of polluting discharges per period where Coast Guard is responsible for providing on-scene coordinator . . . to which such a coordinator is sent;

$x_{11} \equiv$ the number of reported polluting discharges per period to which a Coast Guard representative is sent;

$x_{12} \equiv$ the number of man-hours per period pursuing a public education program;

and where μ_j are obvious upper bounds for these variables (e.g., $\mu_1 =$ the number of transfer operations per period), b_1 is the number of boat hours per period, b_2 is the number of vehicle hours per period, b_3 is the number of personnel hours per period, and a_{ij} is the number of hours of resource i used per unit of activity j for $i=1,2,3$ and $j=1, \ldots, 12$.

In an attempt to solve some of the problems of guidance to field units and to answer some of the questions of marginal utility of additional resources, it was decided to study the setting of priorities for the standards. This was accomplished via preemptive goal programming.

Preemptive goal programming requires the specification of p goals and the partitioning of these goals into K mutually exclusive subsets P_k of $\{1, \ldots, p\}$, where $\alpha \in P_{k'}$, $\beta \in P_{k''}$, and $k' < k''$ if goal α is so important that it must be attained before any resources can be allocated for the (explicit) purpose of attaining goal β. Each subset is called a *priority class*, and within each priority class trade-offs are made via nonpreemptive (for example, weighted linear) goal programming.*

*See Kornbluth [9] and Charnes and Cooper [4] for a survey of goal programming.

In the MEP problem the standards were chosen as goals, so that $p = 12$, and the initial priorities made by the Coast Guard were $P_1 = \{1\}$, $P_2 = \{2,3\}$, $P_3 = \{4,5\}$, $P_4 = \{9,10\}$, $P_5 = \{11\}$, $P_6 = \{6,7,8\}$, and $P_7 = \{12\}$, so that $K = 7$. Given such a priority structure, the determination of activity levels was simply viewed as solving the following preemptive one-sided goal programming problem for $k = 1, \ldots, K$:

$$(2.k) \qquad \text{Min } z_k = \sum_{j \in P_k} w_j \delta_j^-$$

$$\text{subject to } \sum_{j=1}^{12} a_{ij} x_j \leqslant b_i \quad \text{for } i=1,2,3,$$

$$x_j + \delta_j^- - \delta_j^+ = s_j, \quad j=1, \ldots, 12,$$

$$\sum_{j \in P_\eta} w_j \delta_j^- = z_\eta^* \quad \text{for } \eta = 1, \ldots k-1,$$

$$0 \leqslant x_j \leqslant \mu_j \quad \text{for } j=1, \ldots, 12,$$

$$\delta_j^- \geqslant 0 \text{ and } \delta_j^+ \geqslant 0 \quad \text{for } j=1, \ldots, 12,$$

where

$w_j \equiv$ a nonnegative number representing the relative weight of deviations from goal j in priority class P_k for each k;

$z_\eta^* \equiv$ the optimal objective function value to problem $(2.\eta)$;

$s_j \equiv$ the goal (Mission Performance Standard) for activity j (which must be less than or equal to the structural upper bound μ_j);

$\delta_j^- \equiv$ the amount that x_j is below the standard s_j;

$\delta_j^+ \equiv$ the amount that x_j is above the standard s_j;

As noted by others (e.g., Charnes and Cooper [3]) the solution procedure for this problem is a K-stage process completely analogous to the two-phase simplex-method approach. Further, by introducing non-Archimedian weights M_k for each priority class P_k, where $M_k >> M_{k+1}$, the problem can be viewed as a linear program.

Solving the preemptive goal-programming problem (2.k) for the initial priority classes, with the relative weights w_j set equal to unity, we obtained a solution. However, MEP program management did not feel that this solution represented the best utilization of resources and, as a result, made two changes: priority rankings were modified, and threshold levels for specific activities which had zero activity levels were established. These threshold constraints entered the problem (2.k) as lower bounds t_j on the activities, i.e.,

$$(3) \qquad t_j \leqslant x_j \quad \text{for } j=1, \ldots, 12.$$

The preemptive goal-programming problem (2.k) as modified by the threshold constraints (3) was solved with respect to the revised priority structure. However, the new solution was not acceptable either and a second reordering of priorities was given. Again, however, the resulting solution did not appear to make the best utilization of the resources. This led us to evaluate the approach.

2. LIMITATIONS OF GOAL PROGRAMMING

Although, at first, it may seem that priority changes reflect improvements in the understanding by MEP program administrators of their preferences, a closer analysis reveals a basic characteristic of the preemptive goal-programming approach. To understand this characteristic, consider the MEP problem and choose two standards in different priority classes. Then, observe that the level of performance of the two corresponding activities can be characterized by a lexicographical preference ordering between the activities (within the limits of the standards). Now from the argument given in note 2 on pages 72 and 73 of Debreu [6], which shows that lexicographical preferences cannot correspond to a utility function, *we conclude that having a utility function over the twelve activities is inconsistent with the preemptive goal-programming approach to the MEP problem.* (Note that the utility preferences, in this case, would be those of the particular MEP program administrator responsible for the activities and, in particular, his utility function would be his perception of a welfare function* over the activities where, of course, his perception is the result of staff reports and staff opinions.) With respect to this result, the changing of priorities of the standards to get a solution that "represented the best utilization of resources" can be viewed as a futile attempt to model utility preferences in preemptive fashion.

More generally (from analogous reasoning) we have the following result: *for a multiple-objective program with m objectives, having a utility function over the m objectives is theoretically inconsistent with the preemptive goal-programming approach of specifying a goal for each of the m objectives and partioning the goals into different priority classes*†. Note that specifying a goal for each of the m objectives and the partitioning of the goals into different priority classes is the typical approach in applying preemptive goal programming. However, this does *not* mean that the preemptive approach is, in general, incompatible with a utility-preference structure.

Again using the MEP problem as an illustration, we now show how the preemptive approach can be made consistent with a utility-preference structure. More specifically, letting $U(x)$ be a utility function, letting x^* be an optimal solution to the problem

$$\text{(4)} \qquad \text{Max } U(x)$$

Subject to (the same constraints as in (1)),

and letting ϵ be a positive number, we will see how a preemptive approach could theoretically yield a solution $\bar{x}$, where $0 \leqslant U(x^*) - U(\bar{x}) \leqslant \epsilon$. To see this, let $\{c_1, \ldots, c_p\}$ be a set of p positive numbers, where $c_1 < c_2 < \ldots c_p$, where $c_{h+1} - c_h \leqslant \epsilon$ for $h = 1, \ldots, p-1$, and where $c_p \geqslant U(x^*)$. For each $h = 1, \ldots, p$ we define a goal, called goal h, as finding a feasible solution $\hat{x}$ where

$$\text{(5.}h\text{)} \qquad U(\hat{x}) \geqslant c_h$$

or, equivalently,

$$\text{(5.}h\text{)}' \qquad \hat{x} \in \{x: U(x) \geqslant c_h\}.$$

Note that this corresponds to having a solution on or above the indifference (isoutility) contour $\{x: U(x) = c_h\}$. Let $P_h = \{h\}$ for each $h = 1, \ldots, p$ (so that $K = p$). Now, the preemptive

*Of course, as shown by Arrow [1], a true welfare function does not generally exist. We are, however, only considering a perception of one.

†This result does not seem to be generally known — at least, no statement of it was found in the literature.

approach proceeds by finding feasible solutions that satisfy (5.h) for increasingly larger values of h, and this yields solutions with higher and higher utilities. In particular, it yields a feasible solution $\bar{x}$ where, for some h', $c_{h'} \leqslant U(\bar{x}) < c_{h'+1}$ and $c_{h'} \leqslant U(x^*) \leqslant c_{h'+1}$ and, hence, $U(x^*) - U(\bar{x}) \leqslant c_{h'+1} - c_{h'} \leqslant \epsilon$.

Since, in practice, we won't know the functional form of $U(x)$, we might try to approximate the set (5.h)'. If $\tilde{x}$ is a feasible solution where $U(\tilde{x}) = c_h$, then one approximation to (5.h)' is the set $\{x : x \geqslant \tilde{x}\}$. Note that by identifying $\tilde{x}_j$ with the threshold t_j, we see that the Coast Guard, by introducing the threshold constraints (3), was actually modifying the original preemptive goal-programming model in this way. Since changing the priorities of the standards alone cannot yield a good approximation to a utility function, we can view this as an attempt to obtain solutions more consistent with a utility-preference structure.

One possible way to overcome the difficulties of the preemptive approach is simply to abandon it in favor of a nonpreemptive goal-programming approach. One tempting possibility is to simply let $K = 1$ in (2.k) so that we get a weighted linear one-sided goal program (2.1). From an observation by Dyer [7], this corresponds to assuming that

$$U(x) \approx \sum_{j=1}^{12} U_j(x_j),$$

where

$$U_j(x_j) = \begin{cases} U_j(s_j) + w_j(x_j - s_j) & \text{when } x_j \leqslant s_j, \\ U_j(s_j) & \text{when } x_j \geqslant s_j. \end{cases}$$

One problem in using this approach is the well-known difficulty in specifying a priori the goals (s_j) and the deviational weights (w_j). Another problem, not generally recognized, is the tendency for optimal solutions to standard linear goal programs to occur at extreme points. To illustrate this difficulty suppose that standards were chosen such that $s_j \geqslant x_j$ for all x satisfying the constraints to (1). Then, the objective function in (2.1) can be written as a linear function in x so that a solution to (2.1) will almost always (i.e., with probability one) be at a unique extreme point to the constraints in (1). On the other hand, one would not usually expect the activity levels maximizing a utility function to be at an extreme point (especially if the extreme points are not "close" to each other). The above difficulty is symptomatic of a basic problem of standard linear goal programming resulting from using a simple weighted deviation from goals as an approximation to a preference function.

This leads to the desirability of obtaining a better approximation to the utility function. One approach (motivated by this application) is goal-interval programming (GIP) as developed (for this problem) by Charnes, Cooper, Harrald, Karwan, and Wallace [5]. Like standard linear goal programming, this approach implicitly assumes that $U(x) \approx \sum_{j=1}^{12} U_j(x_j)$. However, the GIP approach assumes that each $U_j(x_j)$ is a general piecewise linear function and, in addition, when each $U_j(x_j)$ is concave it shows how the goal program reduces to a linear program. In implementing this approach we must ask how restrictive the assumption that $U(x) \approx \sum_{j=1}^{12} U_j(x_j)$ is. In particular, when can we assume that $U(x)$ is additively separable and how limited is this approach when $U(x)$ is not additively separable?* In the MEP application, for example,

*Subsequent to the submission of this paper, Professor J.S. Dyer independtly presented a paper ("On the Relationship Between Goal Programming and Multiattribute Utility Theory," presented at the International Symposium on Extremal Methods and Systems Analysis, Austin, Texas, September 1977) that discusses these questions of additive separability in goal programming.

certain activities (such as activities 6,7, and 8) are probably intrinsically interrelated so that the additive separability must be questioned. This points the way to investigating other nonlinear functions (including metric based utility (loss) functions discussed in [3]) to obtain better approximations to preferences. In doing so, however, the problem of relating such functions to actual preference data (for example, to estimate the relevant parameters) must be addressed. This means that an integration of methods for quantifying preference data (e.g., the multidimensional scaling methods of Carroll [2] or Srinivasan and Shocker [10]) into the goal programming approach should be made.

REFERENCES

[1] Arrow, K., *Social Choice and Individual Values,* (Wiley, New York, first edition, 1951, and second edition, 1963).

[2] Carroll, J.D., "Individual Differences and Multidimensional Scaling," in *Multidimensional Scaling*, Volume I, edited by R.N. Shepard, A.K. Romney, and S.B. Nerlove (Seminar Press, New York, 1972) pp. 105-155.

[3] Charnes, A., and W.W. Cooper, *Management Models and Industrial Applications of Linear Programming* (Wiley, New York, 1961).

[4] Charnes, A., and W.W. Cooper, "Goal Programming and Multiple Objectives Optimization," European Journal of Operations Research *3*, 347-361 (1976).

[5] Charnes, A., W.W. Cooper, J. Harrald, K. Karwan, and W.A. Wallace, "A Goal Interval Programming Model for Resource Allocation in a Marine Environmental Protection Program," Journal of Environmental Economics and Management *3*, 347-361 (1976).

[6] Debreu, G., *Theory of Value, and Axiomatic Analysis of Economic Equilibrium,* Cowles Foundation Monograph (Yale University Press, New Haven, Connecticut, 1959).

[7] Dyer, J.S., "Internative Goal Programming," Management Science *19*, 62-70 (1972).

[8] Harrald, J., K. Karwan, J. Leotta, and W.A. Wallace, "Accessing the Effectiveness of a Marine Environmental Protection Program," Socio-Economic Planning Sciences (forthcoming).

[9] Kornbluth, J., "A Survey of Goal Programming," Omega, *1*, No. 2 (1973).

[10] Srinivasan, V., and A.D. Shocker, "Linear Programming Techniques for Multidimensional Analysis of Preferences," Psychometrika *38*, 337-369 (1973).

INDEX TO VOLUME 25

CUMULATIVE INDEX FOR VOLUMES 1-25

NEWS AND MEMORANDUM

List of Referees

The Editors of the Naval Research Logistics Quarterly are grateful to the following individuals for assisting in the review of articles prior to publication.

C. J. Ancker, Jr.
L. A. Aroian
H. Ascher
K. R. Baker
V. Balachandran
J. L. Balintfy
R. E. Barlow
D. P. Baron
R. Bartels
Z. Barzily
R. Bechhofer
C. E. Bell
L. Bell
G. K. Bennett
G. E. Bennington
R. Biedenweg
G. Bitran
H. W. Block
J. P. Bonin
J. Bracken
J. Bram
M. Brown
C. Burdet
D. Butler
A. V. Cabot
R. G. Cassidy
M. J. Chandra
S. S. Chitgopekar
R. H. Clark
B. W. Conolly
L. Cooper
R. Cooper
R. Cottle
W. H. Cunningham
C. Das
R. Davis
E. Denardo
D. DePriest
C. Derman
R. L. Disney
J. M. Dobbie
L. Dolansky
J. Donelson, III
M. Dresher
S. D. Dubey
A. Dudek
P. S. Dwyer
R. Ehrhardt
L. Ellwein
H. Emmons
H. Enzer
D. Erlenkotter
J. D. Esary
J. P. Evans
R. Farrell
H. Feingold
R. Feldman
M. Fiorello
M. J. Fischer
P. C. Fishburn
M. L. Fisher
M. Florian
J. O. Flynn
S. Fond
R. L. Francis
H. M. Frank
S. Gass
J. L. Gastwirth
M. A. Geisler
A. E. Gelfand
A. Geoffrion
L. George
N. D. Glassman
F. Glover
A. L. Goel
D. Granot
J. S. Graves
S. C. Graves
P. Gray
H. Greenberg
I. Greenberg
D. Gross
F. E. Grubbs
J. N. D. Gupta
S. C. Gupta
S. S. Gupta
S. Haber
C. M. Harris
J. Haslett
A. Hax
D. P. Heyman
A. J. Hoffman
T. C. Hu
D. A. Hsu
A. P. Hurter, Jr.
D. L. Iglehart
E. Ignall
Y. Ijiri
R. Jagannathan
D. Jansson
R. Jaikumar
D. L. Jaquette
T. Jayachandran
A. J. Kaplan
U. Karmarkar
W. J. Kennedy, Jr.
J. L. Kennington
B. M. Khumawala
W. Kim
M. Klein
P. R. Kleindorfer
D. Klingman
K. O. Kortanek
D. Kreps
C. H. Kriebel
A. Laemmel
L. Lasdon
A. M. Law
L. J. LeBlanc
J. Lehoczky
G. J. Lieberman
S. A. Lippman
R. Love
W. F. Lucas
A. MacCormick
N. R. Mann
W. G. Marchal
C. Marshall
D. Mastran
W. L. Maxwell
G. McCormick
L. McGinnis
P. McKeown
J. Mensching
D. Montgomery
R. C. Morey
J. G. Morris
J. A. Muckstadt
F. H. Murphy
J. M. Myhre
S. Nahmias
P. Nanda
G. Nemhauser
M. F. Neuts
I. Olkin
J. Pang
S. S. Panwalkar
J. H. Patterson

W. Pierskalla
S. R. Pliska
E. L. Porteus
A. J. Quinzi
P. Ranig
D. B. Rosenfield
M. Rosenshine
S. M. Ross
U. Rothblum
M. Rothkopf
H. M. Salkin
S. C. Saunders
L. E. Schrage
L. Schruben
J. Sethuraman
L. Shaw
A. Shogan
E. Silver
N. D. Singpurwalla
D. Smith
R. Soland
Henry Solomon
Herbert Solomon
M. Sovereign
V. Srinivasan
R. M. Stark
D. I. Steinberg
G. R. Stewart
L. D. Stone
P. Sullo
P. Switzer
W. Szwarc
H. A. Taha
A. Tamhane
C. S. Tapiero
J. G. Taylor
R. Thisted
M. Thomas
G. Thompson
W. A. Thompson, Jr.
T. Varley
H. M. Wagner
W. Walker
K. T. Wallenius
E. J. Wegman
R. E. Welsch
R. E. Wendell
A. C. Wheeler
A. Whinston
J. A. White
T. M. Whitin
W. Whitt
W. L. Winston
J. W. Wingate
C. Witzgall
R. Wollmer
C. Wu
O. Yu
S. Zacks
V. A. Zaloom
S. Zionts
A. A. Zoltners

☆ U.S. GOVERNMENT PRINTING OFFICE: 1979 — 281-491/1

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

DECEMBER 1978
VOL. 25, NO. 4
NAVSO P-1278

CONTENTS

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217